Regresión Logística: Fundamentos y aplicación a la investigación sociológica

Texto elaborado por el Equipo Docente
para la asignatura

Análisis Multivariante

Luis Camarero Rioja

Alejandro Almazán Llorente

Beatriz Mañas Ramírez

Departamento de Sociología I, UNED

V.0.2

ÍNDICE

INTRODUCCIÓN

La regresión logística es una técnica analítica que nos permite relacionar funcionalmente una variable dicotómica con un conjunto de variables independientes. El análisis de regresión logística es muy frecuente en muchos campos de investigación, siendo especialmente empleado en investigación socio-sanitaria. Por su capacidad para analizar las relaciones de variables categóricas entre sí tiene una gran importancia en la investigación sociológica. En el análisis de datos sociales, antes que su capacidad para establecer relaciones funcionales y predecir sucesos, su utilidad deriva de la lectura de los coeficientes -Odd Ratio- para interpretar los efectos que tienen las categorías sobre la variable dependiente. Uno de los problemas fundamentales cuando intervienen diversas variables en un fenómeno es determinar cuál es la contribución de cada una de ellas, suponiendo que el resto de las variables no cambian.

Por analogía, la regresión logística puede considerarse una extensión de los modelos de regresión lineal, con la particularidad de que el dominio de salida de la función está acotado al intervalo [0,1] y que el procedimiento de estimación, en lugar de mínimos cuadrados, utiliza el procedimiento de estimación máximo-verosímil. Para el ajuste de modelos de regresión logística resulta esencial el empleo de programas informáticos.

Este capítulo tiene como objeto presentar el análisis de regresión logística desde el uso y aplicaciones de la investigación sociológica. Para ello se centra en un ejemplo sencillo con datos reales: el análisis de la participación en huelgas. En primer lugar se presentan los fundamentos, especialmente la noción de Odd o razón y el Logit. En un segundo momento se comienza poniendo en práctica el modelo con variables sencillas, para ir complejizando el tipo y número de variables mientras se muestran los principales estadísticos que ofrecen los programas de ordenador. Por último se hace una discusión y ajuste del modelo.

I. HERRAMIENTAS CONCEPTUALES Odd y Logit.

1. Razón, Odd y Odd Ratio

Observemos la siguiente tabla de contingencia que señala la participación en una huelga[1] según sexo. (Estudio CIS 2941-Abril 2012).

[1] Sólo aquéllos que han declarado haber participado en una huelga durante los últimos 12 meses. El universo, como es habitual en encuestas de opinión, se refiere a los mayores de 18 años.

Tabla 1. Participación en una huelga por sexo. Valores absolutos

	Hombre	Mujer	Total
No han participado	960	1057	2017
Han participado	261	206	468
Total	1221	1263	2484

La tabla muestra que hay una mayoría que no ha participado en una huelga. En concreto, hay 468 entrevistados que han participado de un total de 2484. En porcentajes, tenemos que el 18,8% participa frente a un 81,2% que no participa. Aunque estamos acostumbrados a expresar los datos en proporción o porcentajes, podríamos hacerlo también mediante una razón. Por ejemplo, podemos señalar que (468/2017=0,23) hay 0,23 participantes por cada uno que no participa, o lo que es equivalente: 23 participantes por cada 100 que no participan. También podemos expresar la lectura inversa (2017/468=4,31): hay 4,3 que no participan por cada uno que participa, o bien 431 que no participan por cada 100 que participan.

Una **razón** o **ratio** es el cociente entre dos cantidades y señala cuantas veces una cantidad es mayor o menor respecto a la otra.

La lectura "clásica" que hacemos de la tabla anterior (tabla 1) se realiza mediante el uso de porcentajes. La variable dependiente es la *participación* y la variable independiente el *sexo*. Los porcentajes en dirección de la variable independiente (en este caso en columnas):

Tabla 2. Participación en una huelga por sexo. Porcentajes verticales

	Hombre	Mujer	Total
No han participado	78,6%	83,7%	81,2%
Han participado	21,4%	16,3%	18,8%
	100%	100%	100%

Los datos nos indican que los hombres participan más en huelgas que las mujeres. ¿Cuánto más? Podemos calcular la diferencia de porcentajes: 21,4%-16,3%=5,1%. Los hombres participan un 5,1% más en huelgas que las mujeres.

Pero también podemos observar la tabla anterior en términos de razón. Por ejemplo, el 21,4% de los hombres declaran haber hecho huelga mientras que el 78,6% restante declara no haber participado. La relación entre hombres huelguistas y no huelguistas es: 21,4/78,6=0,272. La ratio la leemos de la siguiente forma: hay 0,272 hombres que hacen huelga por cada uno que no la hace, de manera aproximada podemos decir que la

relación[2] en los hombres entre hacer y no hacer huelga es de 3 a 11. De forma habitual esta razón o ratio suele denominarse con el término Odd.

$$Odd = \frac{p}{q} = \frac{p}{(1-p)}$$

El término Odd en inglés se refiere a la razón que se establece entre la ocurrencia -o su probabilidad- de un suceso respecto a su no ocurrencia. Se interpreta como ventaja comparativa. Es muy usual, por ejemplo, en el mundo de las apuestas.

Procediendo de la misma forma podemos concluir que hay 0,195 mujeres que hacen huelga por cada una que no la hace (de forma aproximada 1 a 5).

Podemos interpretar el Odd en términos de probabilidad. En los casos anteriores podemos señalar que la probabilidad de encontrar aleatoriamente una mujer que hace huelga es la quinta parte respecto a la de encontrar una mujer que no hace huelga.

	Hombre	Mujer
$Odd_{Participa/No\ participa}$	0,272	0,195

Como podemos ver, a través de las ratio llegamos a la conclusión de que los hombres participan más en huelgas que las mujeres. Para responder a la pregunta "¿cuánto más participan los hombres que las mujeres?" podemos utilizar también un ratio, procediendo así:

Mujeres/Hombres = 0,195/0,272 = 0,717

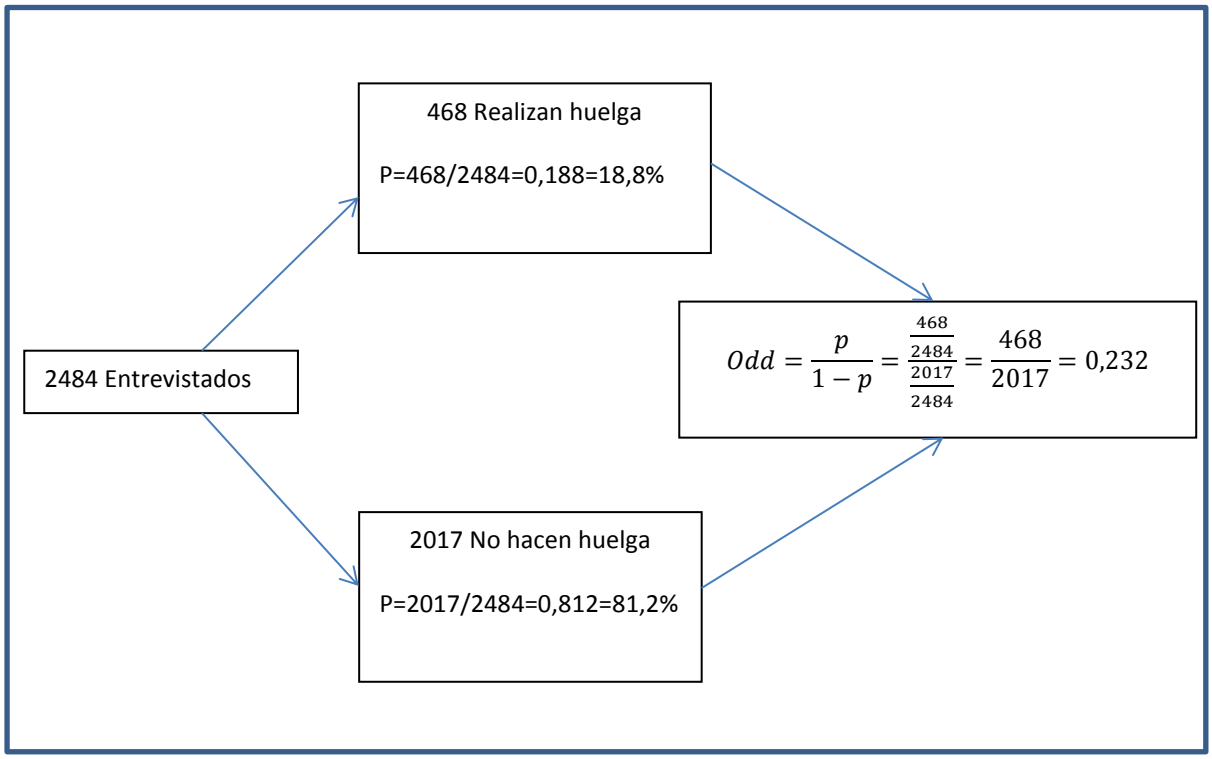
Este ratio de ratios, lo leemos así: la probabilidad de encontrar una mujer que hace huelga sobre una que no la hace es de 0,717 veces respecto al caso de los varones. La probabilidad se reduce, por tanto, un 28,3% respecto a la de los hombres.

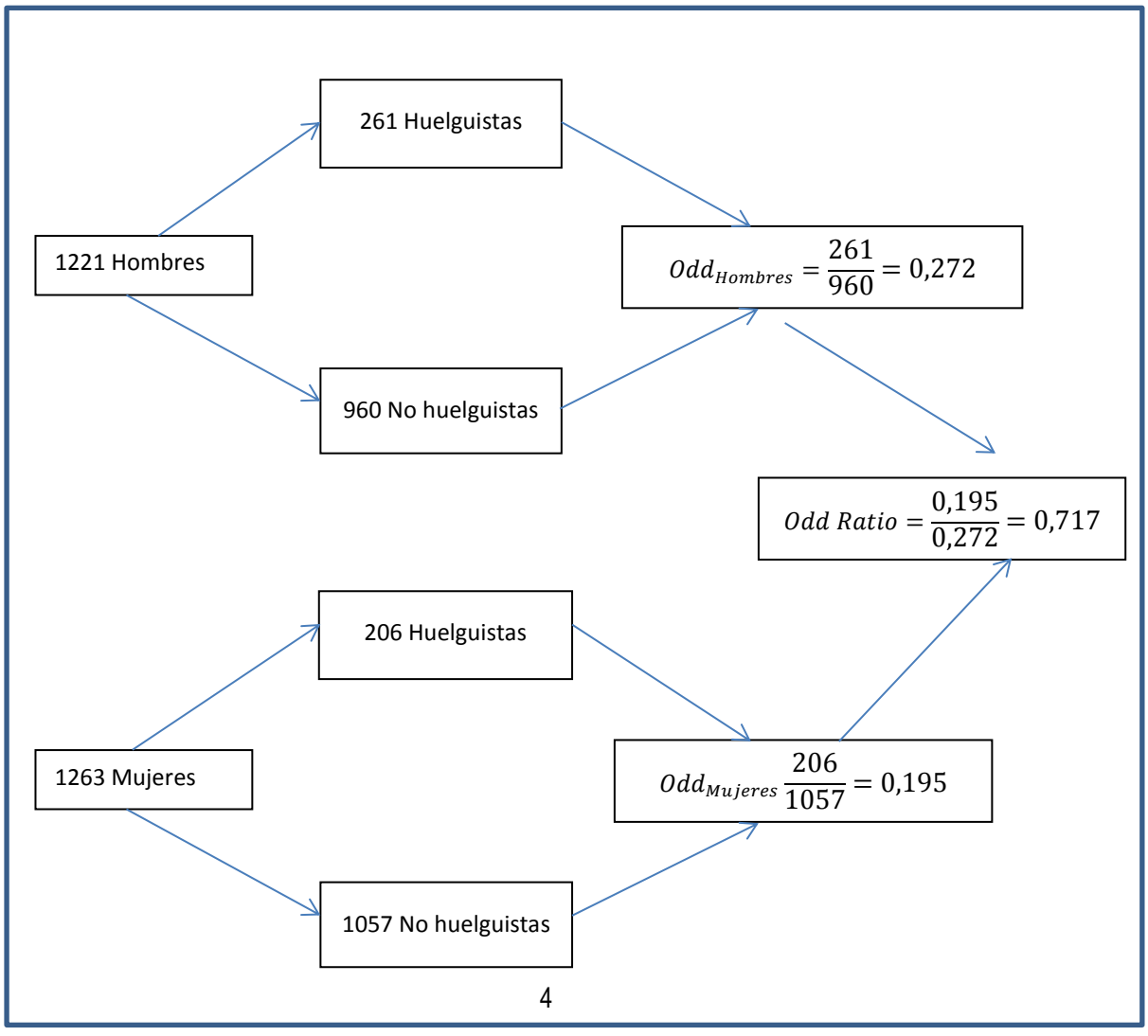
Este término, que es una razón de Odds, se denomina Odd Ratio -abreviadamente OR- y puede interpretarse como ventaja comparativa o, como razón de probabilidades. En nuestro caso, la probabilidad de encontrar entre las mujeres una que haga huelga es menor que en el caso de los hombres. Cuando el Odd Ratio alcanza el valor 1 quiere decir que no hay diferencias.

$$Odd\ Ratio = OR = \frac{Odd_A}{Odd_B} = \frac{\frac{p_A}{(1-p_A)}}{\frac{p_B}{(1-p_B)}}$$

[2] Obsérvese que 3 a 11 es 3/11=0,273 valor próximo a 0,27.

Vamos a ordenar esquemáticamente la información:





2. La relación entre Odd y proporción. El Logit

A partir de la noción de Odd podemos ir pensando en la forma de mostrar una relación algebraica que nos permita indicar la probabilidad de respuesta -afirmativa- que tiene un entrevistado escogido al azar.

En primer lugar vamos a mostrar la relación que existe entre Odd Ratio y Proporción -entendida como probabilidad-. Anteriormente hemos definido que

$$Odd = \frac{p}{1-p}$$

Y mediante dicha definición la relación entre Odd y p podemos expresarla de la siguiente forma:

$$p = \frac{Odd}{(1+Odd)}$$

Si partimos de $\mathrm{Odd} = \frac{\mathrm{p}}{1-\mathrm{p}}$, mediante sencillas transformaciones algebraicas podemos mostrar que:

$$Odd = \frac{p}{1-p}$$

$$(1-p)Odd = p$$

$$Odd - p \cdot Odd = p$$

$$Odd = p + p \cdot Odd$$

$$Odd = p(1+Odd)$$

$$\frac{Odd}{(1+Odd)} = p$$

Como podemos observar, los odds varían desde desde 0 a +∞. Téngase en cuenta que p varía desde 0 a 1. Por lo tanto, cuando p está muy cerca de 1:

$$Odd = \frac{p}{1-p} = \frac{1}{0} = +\infty$$

Y en el caso inverso:

$$Odd = \frac{p}{1-p} = \frac{0}{1} = 0$$

A partir del Odd podemos definir el logit, simplemente como el logaritmo[3] del Odd. Así:

$$\text{Logit}=\text{Ln}(\text{Odd})=\text{Ln}\left(\frac{p}{1-p}\right)$$

El Logit tiene dos propiedades que nos serán muy útiles, por una parte puede tomar cualquier valor real entre -∞ y +∞. Por otra parte permite una lectura simétrica de la relación entre proporciones.

Por ejemplo, vamos a considerar una población en la que el 30% son hombres y el 70% son mujeres. Podemos definir p, como la proporción de hombres. Entonces, en este caso, el Odd sería 0,3/0,7=0,429

Si, por el contrario, definimos p como la proporción de mujeres obtendríamos que el Odd sería 0,7/0,3=2,333

Es evidente que optar por la proporción o el complemento de la misma ofrece dificultades claras de interpretación. Cuando p es menor que q, el Odd se moverá entre 0 y 1, mientras que cuando p>q el Odd se moverá entre 1 y +∞. Sin embargo tenemos el mismo motivo para considerar tanto la característica como su complementario, tal como ocurre en una población entre hombres o mujeres.

Sin embargo, mediante la transformación logarítmica del Odd los valores son comparables sin importar cuál sea la categoría tomada para el cálculo de la proporción. Así, el logit para la población 30% hombres y 70% mujeres:

Odd=0,3/0,7=0,429 Logit=Ln(0,3/0,7)=-0,847

Odd=0,7/0,3=2,333 Logit=Ln(0,7/0,3)=+0,847

Como podemos observar el valor absoluto es el mismo con independencia de la categoría elegida para definir p o q. La única variación está en el signo. Este nos indica si es positivo que p>q o la situación contraria cuando es negativo.

[3] En regresión logística se emplea siempre el logaritmo neperiano o de base "e".

Gráfico 1

Odds según valores de *p*

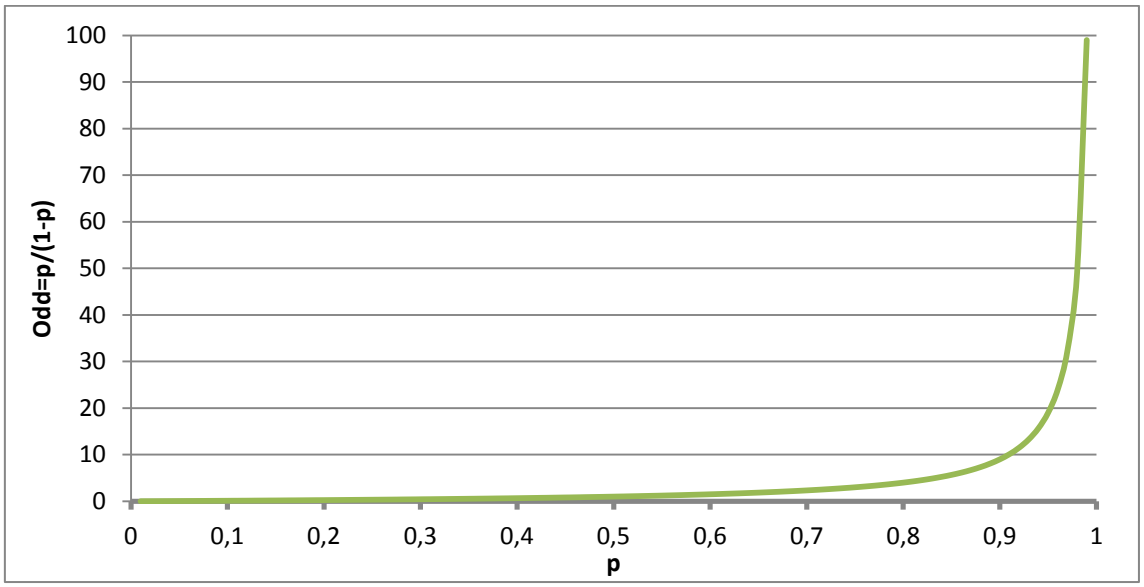
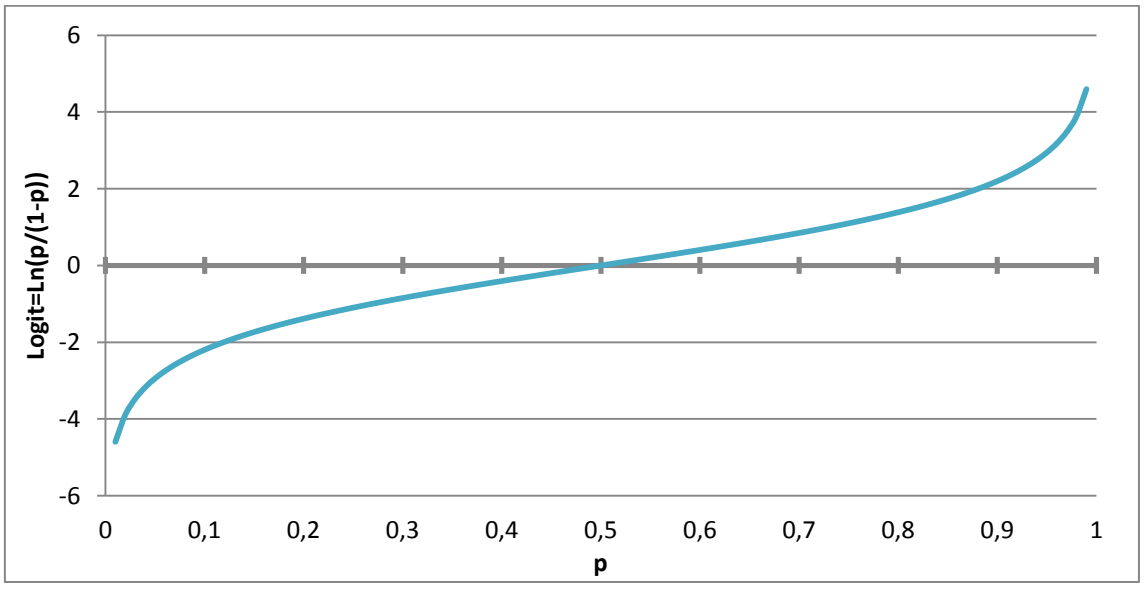


Gráfico 2.

Logit según valores de *p*



También podemos definir el Logit para un Odd Ratio. Si el Odd Ratio entre dos categorías -por ejemplo, hombres y mujeres- lo definimos como:

$$Odd\ Ratio = \frac{\frac{p_h}{q_h}}{\frac{p_m}{q_m}}$$

El logit será:

$$Ln\left(\frac{\frac{p_h}{q_h}}{\frac{p_m}{q_m}}\right) = Ln\left(\frac{p_h}{q_h}\right) - Ln\left(\frac{p_m}{q_m}\right)$$

Es decir, el logit del Odd ratio es la diferencia entre los logit de los Odds.

3. La construcción de una función Logit

Nuestro interés es la elaboración de un modelo que nos permita determinar "p", la probabilidad. Cuando hablamos de probabilidad la entendemos de forma genérica como la posibilidad de ocurrencia de un suceso que puede consistir en la posesión de una característica -por ejemplo, ser propietario de una vivienda, o de un empleo- o, como sucede en el estudio de opiniones, de una respuesta afirmativa o negativa respecto a un ítem.

Como p varía de 0 a 1, hemos definido el Logit, que es otra forma de expresar p pero que, como vamos a mostrar, nos permite hallar un camino para encontrar una relación algebraica entre una probabilidad y un conjunto de variables de categoría.

Como el Logit puede tomar cualquier valor real (desde -∞ a +∞), podemos suponer un modelo lineal sin restricciones[4]. Si llamamos z al logit:

$$z=\alpha+\beta x$$

$$\text{Ln}\left(\frac{p}{1-p}\right)=\alpha+\beta x$$

Vemos que α+βx define una recta donde "x" será el valor que toma la variable explicativa. El coeficiente α, al igual que la constante en la recta de regresión, supone el valor que toma la variable independiente cuando la variable dependiente toma el valor 0 -es decir, si la variable independiente no tiene efecto- mientras que β, al igual que la pendiente de la recta de regresión, supone el incremento en valor de la variable dependiente cuando la independiente crece en una unidad.

Si en la expresión anterior

$$\text{Ln}\left(\frac{\text{p}}{1-\text{p}}\right)=\alpha+\beta x$$

[4] Si intentáramos ajustar un modelo p=α+βx, deberíamos poner restricciones a los coeficientes α y β para que p se mantuviera entre 0 y 1, dentro de un intervalo acotado de valores de x.

quitamos el logaritmo, obtenemos que:

$\left(\frac{p}{1-p}\right)=e^{(\alpha+\beta x)}$

Y si despejamos p, podemos expresar el modelo como:

$$p = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}}$$

O también como:

$$p = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

O considerando z como el Logit: ($z= \alpha + \beta x$)

$$p = \frac{e^{z}}{1+e^{z}} \qquad\qquad p = \frac{1}{1+e^{-z}}$$

Téngase en cuenta también, haciendo uso de las relaciones exponenciales, que:

$$e^{\alpha+\beta x} = e^{\alpha}e^{\beta x}$$

Si bien, como se verá más adelante, podemos emplear una variable de categoría como independiente, vamos a desarrollar el modelo pensando en una variable dicotómica. En nuestro caso, queremos ajustar la tabla 1 y utilizaremos como variable independiente el sexo. Dicha variable la codificamos de la siguiente forma: 0 cuando el caso es hombre y 1 cuando es mujer. Esta codificación de "ceros" y "unos" se denomina "dummy". Será explicada en detalle más adelante cuando se utilice una variable con más de dos categorías.

A partir de los datos, de la tabla 1, expresados en proporciones, podemos calcular los Logit.

Tabla 3. Participación en una huelga por sexo. Datos expresados en proporciones

	Hombre	Mujer	Total
No Huelguista	0,78624079	0,83689628	0,81199678
Huelguista	0,21375921	0,16310372	0,1884058
	1	1	1

El Logit cuando X=0 (es decir, para hombres):

$$\text{Logit}=\text{Ln}\left(\frac{0,21375921}{0,78624079}\right)=\text{Ln}\,(0,271875)=-1,30241288$$

Cuando X=1 (para mujeres):

$$\text{Logit}=\text{Ln}\left(\frac{0,16310372}{0,83689628}\right)=\text{Ln}\,(0,1948912)=-1,63531382$$

En el caso de una variable independiente dicotómica, el ajuste de la función a partir de los Logit resulta muy sencillo.

Cuando X=0 (hombres), entonces $\beta x=\beta(0)=0$, luego Logit=α

Por lo tanto α= -1,30241288, que es el valor del Logit (para X=0)

Y cuando X=1 (mujeres), una vez conocido α, simplemente:

-1,63531382= -1,30241288+β(1)

Luego: β=-1,63531382+1,30241288=-0,33290094

Así: $\alpha+\beta x$=-1,302-0,333x

Observemos que una vez conocidos α y β, podemos calcular las probabilidades, cuando X=0 (hombres)

$$\text{p}=\frac{1}{1+\text{e}^{-(\alpha+\beta x)}}=\frac{1}{1+\text{e}^{-(-1,302)}}=\frac{1}{1+3,678}=0,214$$

21,4% es, cómo podemos comprobar, el porcentaje de hombres que realizan huelga. Y para mujeres:

$$\text{p}=\frac{1}{1+\text{e}^{-(\alpha+\beta x)}}=\frac{1}{1+\text{e}^{-(-1,302-0,333)}}=\frac{1}{1+5,129}=0,163$$

16,3% es el porcentaje de mujeres que realizan huelga.

Para interpretar correctamente los coeficientes vamos a observar las siguientes propiedades:

La constante α=-1,302 $e^{\alpha} = e^{-1,302} = 0,272 = Odd_{hombres}$

La pendiente β=0,642 $e^{\beta} = e^{-0,333} = 0,717 = OR$

Cuando estamos ante variables independientes de categoría, los términos "efectos" y "relación" tienen una interpretación ligeramente distinta que en el caso de las variables de intervalo. Observemos que el modelo Logit se construye, no en términos absolutos, sino a través de la comparación mediante el uso de una razón. El Logit nos dice "cuánto más respecto a...". Dicho de otra forma, utilizamos una categoría como patrón de medida -respecto a. El Logit o el Odd, como razones, señalan cuántos elementos poseen una característica respecto a los que no la tienen -estos últimos son empleados como patrón.

Los coeficientes β son, como hemos visto anteriormente, Odds Ratio. Es decir relacionan una categoría respecto a otra. En el caso concreto que estamos analizando, en el que hablamos de mujeres por cada hombre, los hombres son la categoría que sirve de comparación. Como el valor de OR<1 quiere decir que hay menos mujeres que hombres que hacen huelga, en concreto la relación es de 0,717, que podemos expresar como: "hay 72 mujeres que hacen huelga por cada cien hombres que también hacen huelga".

En algunas interpretaciones del coeficiente β podemos pensar en términos de ganancia o de pérdida en el supuesto de cambio de categoría. Difícilmente podemos cambiar de sexo, pero podríamos cambiar de trabajo, o de lugar de residencia. En nuestro caso, podríamos decir que en el paso de hombre a mujer la probabilidad de hacer huelga se reduce en un 28,3% (1-0,717=0,283). O también, de forma equivalente en el paso de mujer a hombre la probabilidad de hacer huelga aumenta un 71,7%.

Por lo general, como veremos en los siguientes ejemplos, la interpretación descansa sobre los coeficientes e^{β}. El término α (o constante) no suele tener un significado de interés cuando hay más de una variable independiente. Podría considerarse como el efecto cuando todas las variables independientes valen 0.

En este caso, que sólo hay una variable independiente, el término α nos indica cuál sería la relación cuando no interviene la variable independiente, es decir si todos los elementos fueran iguales. En este caso estamos suponiendo que todos son hombres.

II. LA GENERALIZACIÓN DE UN MODELO Logit

La ventaja de uso de los modelos Logit es el análisis conjunto de un grupo de variables independientes. Como veremos, los coeficientes OR muestran el efecto que tiene una categoría sobre la variable dependiente, bajo el supuesto de que todas las demás variables permanecen constantes.

Observemos ahora la distribución de la respuesta a la pregunta de participación en huelas por la variable nivel de estudios.

Tabla 4. Participación en una huelga por nivel de estudios. Datos absolutos

	Nivel de Estudios		
	Hasta Básicos	Medios	Universitarios
No hacen huelga	1116	554	344
Sí hacen huelga	138	182	145
Total	1254	736	489

Tabla 5. Participación en una huelga por nivel de estudios. Porcentajes verticales

	Nivel de Estudios		
	Hasta Básicos	Medios	Universitarios
No hacen huelga	89,0%	75,3%	70,3%
Sí hacen huelga	11,0%	24,7%	29,7%
Total	100%	100%	100%

De forma rápida podemos comprobar que hay una mayor proporción de seguimiento de jornadas de huelga cuanto mayor es el nivel de estudios. Además de porcentajes, podemos expresar los datos mediante Odd Ratio. Tomamos como referencia la categoría de estudios "hasta básicos" -la que aparece en primer lugar-, y obtenemos los siguientes OR:

$$OR_{Medios/Básicos} = \frac{\frac{182}{554}}{\frac{138}{1116}} = 2{,}657$$

$$OR_{Universitarios/Básicos} = \frac{\frac{145}{344}}{\frac{138}{1116}} = 3{,}409$$

Quienes tienen estudios medios realizan huelgas 2,657 veces más respecto a quienes tienen menores niveles de estudios, y quienes han realizado estudios universitarios la relación es de 3,4 veces mayor.

1. Codificación de variables Dummy

Para introducir esta variable, que tiene tres categorías, en el modelo debemos recurrir a la categorización "dummy". La categorización dummy consiste en la generación de variables dicotómicas para las distintas categorías de la variable. Estas nuevas variables se denominan *ficticias*.

Habíamos codificado la variable sexo, que tiene dos categorías, con 0 cuando era hombre, y 1 cuando era mujer. Se trata de una codificación binaria. Y se realiza con independencia de los códigos alfanuméricos utilizados en el cuestionario. Las encuestas del CIS utilizan 1 para hombres y 2 para mujeres. El INE, para evitar errores de grabación, utiliza 1 para hombres y 6 para mujeres. En la codificación binaria, 0 quiere decir que no se posee la característica y 1 que sí se posee.

En el caso de variables de dos categorías, la codificación binaria resulta evidente. Cuando hay más categorías es también sencillo aunque un poco más laborioso. Con dos categorías, una de las mismas puede considerarse como la característica y la otra como la ausencia de ella. En el caso del sexo, la característica se definió como "ser mujer" (1) y la ausencia de característica como "no ser mujer" (0), que es la característica complementaria.

Para la variable nivel de estudios, si definimos una de sus categorías como característica, la ausencia de esta no es una categoría sino un conjunto de categorías. Por ejemplo, si consideramos que la característica es "tener estudios universitarios", la ausencia de la característica la cumplen quienes tienen estudios medios e inferiores. Como podemos ver en el cuadro adjunto hay varias posibilidades:

Hasta Básicos	Medios	Universitarios
1	0	0
0	1	0
0	0	1

Es decir, podemos construir tres variables dicotómicas:

a) Tienen hasta estudios básicos (1). Tienen más de estudios básicos (0)
b) Tienen estudios medios (1). Tienen estudios superiores o inferiores a los medios (0)
c) Tienen estudios universitarios (1). No tienen estudios universitarios (0).

Con la codificación "dummy", la tabla 4 puede descomponerse en un conjunto de 3 tablas:

Tabla 6. Participación en una huelga por nivel de estudios como variable ficticia. Datos absolutos y porcentajes verticales.

	Hasta Básicos	
	Sí (1)	No (0)
No hacen huelga	1116	898
Sí hacen huelga	138	327
Total	1254	1225

	Hasta Básicos	
	Sí (1)	No (0)
No hacen huelga	89,0	73,3
Sí hacen huelga	11,0	26,7
Total	100%	100%

	Medios	
	Sí (1)	No (0)
No hacen huelga	554	1460
Sí hacen huelga	182	283
Total	736	1743

	Medios	
	Sí (1)	No (0)
No hacen huelga	75,3	83,8
Sí hacen huelga	24,7	16,2
Total	100%	100%

	Universitarios	
	Sí (1)	No (0)
No hacen huelga	344	1670
Sí hacen huelga	145	320
Total	489	1990

	Universitarios	
	Sí (1)	No (0)
No hacen huelga	70,3	83,9
Sí hacen huelga	29,7	16,1
Total	100%	100%

Realmente, de una variable con tres categorías hemos construido tres variables dicotómicas. Sin embargo, no son variables independientes, en la medida en que se solapan. Si volvemos al caso de la variable sexo, también podemos construir dos variables dicotómicas: Mujeres [sí(1)/no(0)] y Hombres [sí(1)/no(0)]. Sin embargo, con una sola de ellas podemos determinar el modelo. Así lo hicimos, sólo empleamos un coeficiente β. Como veremos a continuación para cada variable de categoría, vamos a necesitar k-1 coeficientes (OR) para ajustar el modelo (k es el número de categorías de la variable). Para la variable sexo, como k=2, necesitamos sólo un coeficiente; para nivel de estudios de 3 categorías emplearemos 2 coeficientes.

2. El ajuste con una variable

Para introducir el uso de la regresión logística a diferentes variables comenzaremos ajustando una sola variable de más de dos categorías, como es el nivel de estudios, y posteriormente añadiremos la variable sexo. Para terminar la generalización se introducirá una tercera variable, pero de carácter continuo: la variable edad.

En el apartado anterior, para la variable sexo habíamos utilizado la siguiente relación:

$Z=\alpha+\beta x=-1{,}302-0{,}333x$ siendo los valores de x=0 (hombres) y x=1(mujeres)

En el caso del nivel de estudios, con tres categorías, la ecuación será:

$$Z=\alpha+\beta_1 x_1+\beta_2 x_2$$

A partir de aquí, en esta ecuación sustituiremos las letras griegas por letras latinas. Es decir:

$$Z=a+b_1 x_1+b_2 x_2$$

Como recordará de los cursos de estadística, las letras griegas hacen referencia al parámetro o valor poblacional, mientras que las letras latinas se refieren al estadístico obtenido de una distribución muestral. La constante α la denominaremos "a" y los coeficientes β los representamos como "b". (En algunos textos, la constante se representa con el término b_o: $Z=b_o+b_1 x_1+b_2 x_{2)}$

Para ajustar la variable "nivel de estudios" necesitamos dos coeficientes porque tenemos dos OR que determinar. Aunque son tres categorías, al utilizar una como referencia sólo hay dos OR independientes[5]:

$OR_{Medios/Básicos}$ y $OR_{Universitarios/Básicos}$

El esquema de codificación que utilizaremos, teniendo en cuenta que la categoría de referencia es "hasta básicos", será el siguiente:

	X_1	X_2
Hasta Básicos	0	0
Medios	1	0
Universitarios	0	1

Tomando como referencia la primera categoría de estudios -"hasta básicos"- vamos a calcular los coeficientes en función de los OR.

La constante α es el Odd de la categoría de referencia:

$a=Ln(Odd_{básicos})=Ln(0,11/0,89)=-2,09$

b_1 será el $Ln(OR_{medios/básicos})=Ln(2,657)=0,977$

b_2 será el $Ln(OR_{universitarios/básicos})=Ln(3,409)=1,226$

De esta forma:

[5] Tal vez haya pensado que se puede determinar también el OR que relacione "medios" con "universitarios". Sin embargo, compruebe que $OR_{m/u}=OR_{m/b}-OR_{u/b}$. Es decir, con tres elementos sólo hay dos relaciones independientes. De forma genérica k-1.

	$Z=a+b_1x_1+b_2x_2$	$p=\frac{1}{1+e^{-z}}$
Hasta Básicos	z=-2,09+0,977(0)+1,226(0)=-2,09	$p=\frac{1}{1+e^{2,09}}=0,11$
Medios	z=-2,09+0,977(1)+1,226(0)=-1,113	$p=\frac{1}{1+e^{1,113}}=0,2473$
Universitarios	z=-2,09+0,977(0)+1,226(1)=-0,864	$p=\frac{1}{1+e^{0,864}}=0,2965$

Como puede observarse, la ecuación:

$$p=\frac{1}{1+e^{-(a+b_1x_1+b_2x_2)}}$$

nos permite reconstruir la tabla de porcentajes. Vamos a continuación a presentar el ajuste con dos variables independientes.

3. El ajuste con 2 variables

La tabla siguiente nos permite observar de forma conjunta el efecto de sexo y nivel de estudios sobre la participación en huelgas:

Tabla 7. Participación en huelgas según sexo y nivel de estudios. Absolutos y proporción.

Estudios	Sexo	No han realizado huelga	Sí han realizado huelga	Total	P (Han realizado huelga)
Hasta Básicos	Hombre	519	79	598	0,13210702
	Mujer	597	59	656	0,08993902
Medios	Hombre	282	103	385	0,26753247
	Mujer	272	79	351	0,22507123
Universitarios	Hombre	158	78	236	0,33050847
	Mujer	186	67	253	0,26482213

El cálculo de los coeficientes cuando hay más variables resulta complejo y se utilizan distintos algoritmos y rutinas por aproximación. El programa de ordenador utilizado nos proporciona los siguientes resultados:

	B	Exp(B)
Estudios (1)	0,966	2,627
Estudios (2)	1,229	3,419
Sexo (1)	-0,322	0,724
Constante	-1,932	0,145

Tenemos el coeficiente B así como el valor del exponente del coeficiente. Obsérvese que $Exp(B)=e^B$

Por ejemplo, $e^{0,966}=2,627$

Vamos a examinar por separado el significado de ambos valores. En primer lugar, los coeficientes que nos permiten construir la ecuación para determinar la probabilidad en cada categoría.

En este caso, como tenemos dos variables, una con dos categorías y la otra con tres, la ecuación tendrá la siguiente forma:

$Y=a_o+b_1(Sexo1)+b_2(Estudios1)+b_3(Estudios2)$

Recordemos que hemos codificado las variables de la siguiente forma:

Sexo	**Estudios**
Hombres=0	Hasta Básicos (0,0,0)
Mujeres=1	Medios (0,1,0)
	Superiores (0,0,1)

La asignación de los coeficientes es la siguiente:

Estudios	Sexo	Constante	Sexo(1)	Estudio(1)	Estudio(2)
Hasta Básicos	Hombre	X			
	Mujer	X	X		
Medios	Hombre	X		X	
	Mujer	X	X	X	
Universitarios	Hombre	X			X
	Mujer	X	X		X

Los valores de nuestra ecuación serán:

Y=1,932-0,322(Mujeres)+0,966(Medios)+1,229(Universitarios)

Si desarrollamos las ecuaciones obtenemos que:

Estudios	Sexo		Z	$p=\frac{1}{1+e^{-z}}$	P (Han realizado huelga)
Hasta Básicos	Hombre	Z=-1,932	6,90330305	0,12652938	0,13210702
	Mujer	Z=-1,932-0,322	9,52576278	0,09500499	0,08993902
Medios	Hombre	Z=-1,932+0,966	2,62741376	0,2756785	0,26753247
	Mujer	Z=-1,932-0,322+0,966	3,62552824	0,21619152	0,22507123
Universitarios	Hombre	Z=-1,932+1,229	2,01980304	0,33114743	0,33050847
	Mujer	Z=-1,932-0,322+1,229	2,78709546	0,26405461	0,26482213

4. Significación de los coeficientes

Vamos a observar ahora las salidas directas de ordenador para el análisis conjunto de sexo y nivel de estudios.

	B	E.T.	Wald	gl	Sig.	Exp(B)
SEXO(1)	-,322	,106	9,279	1	,002	,724
ESTUDIOS			98,031	2	,000	
ESTUDIOS(1)	,966	,125	60,156	1	,000	2,627
ESTUDIOS(2)	1,229	,134	83,836	1	,000	3,419
Constante	-1,932	,103	354,159	1	,000	,145

Como podemos observar, los coeficientes vienen acompañados de distintos estadísticos, que vamos a examinar ahora en detalle. En la primera columna, B, tenemos los coeficientes, que nos permiten calcular z=a+bx. En la última columna encontramos los valores exponenciales de los mismos ($0{,}724=e^{-0{,}322}$). Como hemos visto, Exp(B) son los Odd Ratio (OR).

En la segunda columna, encontramos E.T, que es la abreviatura de Error Típico. Como se recodará, el error típico es la desviación típica de la distribución muestral de un estadístico. A partir del error típico podemos determinar el intervalo de confianza del coeficiente.

Por ejemplo, en nuestro caso, b=-0,322 y $E.T. = \sigma_{\hat{b}} = 0{,}106$, podemos construir un intervalo al 99,7% de confianza (Z=3):

$$B \pm Z\sigma_{\hat{b}}$$

$$-0{,}322 \pm 3x0{,}106$$

$$-0{,}322 \pm 0{,}318 = [-0{,}64 : -0{,}004]$$

Como podemos ver en el intervalo de b, con un nivel de confianza mayor del 99,7%, no se encuentra el valor 0. De la misma forma el intervalo del OR, para el mismo nivel de confianza sería:

	b	e^b
Límite inferior	-0,64	0,527
	-0,322	0,724
Límite superior	-0,004	0,996

Con una probabilidad muy alta, ni b para la variable de sexo puede ser nulo, ni el OR de mujeres sobre hombres puede ser 1. Recuerde que un OR=1 significaba que no había efecto. En este caso, podemos asegurar que el sexo tiene efecto. A continuación mostraremos que podemos rechazar la hipótesis nula que indica que b=0 o que $e^b=1$

para un nivel de confianza del 99,98%, lo que equivale a decir que tenemos que considerar el efecto del sexo en la participación en manifestaciones[6].

La distribución muestral de los coeficientes b en el muestreo no es evidente. El cálculo del error típico se hace mediante procedimientos asintóticos. Por todo ello, para contrastar la hipótesis de que b=0 o que e^b=1, se utiliza el estadístico de Wald. En lo que respecta a nuestros intereses simplemente necesitamos saber que el estadístico de Wald se interpreta como una distribución Ji-cuadrado. En función de los grados de libertad, en la siguiente columna nos indica el nivel de significación de cada coeficiente.

El estadístico de Wald se determina mediante la relación entre la diferencia del valor estimado de b respecto al valor de contraste (b=0) y el error típico. La expresión algebraica:

$$Wald = \frac{\left(\hat{b} - b_0\right)^2}{(ET_{\hat{b}})^2}$$

Por ejemplo, para la variable sexo, tomará el valor:

$$Wald_{sexo(1)} = \frac{\left(\hat{b} - 0\right)^2}{(ET_{\hat{b}})^2} = \frac{(-0{,}322 - 0)^2}{0{,}106^2} = 9{,}2$$

Contraste el valor[7] 9,2 en cualquier tabla de Ji-cuadrado, para la fila de 1 grado de libertad. Observará que el nivel de significación está muy cerca de 0,0025. (Concretamente, para ese nivel de significación (ns) y grados de libertad (gl), el valor Jicuadrado=9,404)

Como vemos en el caso de SEXO, la significación es 2 por mil. Esto quiere decir que podemos rechazar la hipótesis nula, b=0 ó que e^b=1, con una probabilidad de equivocarnos menor a 2 por mil. En el caso de los coeficientes de ESTUDIO o de la constante, la probabilidad de equivocarnos si indicamos que tienen efecto en la variable dependiente es menor de 1 por 10.000.

Es decir, nuestro modelo, dice que en la población sexo y nivel de estudios tienen efecto en la participación en huelgas.

[6] Si bien todos los datos hasta ahora muestran la relación entre sexo y participación en huelgas, se podrá comprobar más adelante, cuando especifiquemos un modelo de participación, como dicho efecto no es real, sino espurio.

[7] Hay una ligera diferencia entre el Wald que hemos calculado y el que indica el programa debido a que en nuestros cálculos hemos trabajado únicamente con tres decimales).

5. Lectura de coeficientes

En términos prácticos de la hoja de resultados que nos ofrecen los programas informáticos, debemos leer tres columnas b, e^b y nivel de significación:

	B	Sig.	Exp(B)
SEXO(1)	-,322	,002	,724
ESTUDIOS		,000	
ESTUDIOS(1)	,966	,000	2,627
ESTUDIOS(2)	1,229	,000	3,419
Constante	-1,932	,000	,145

En buena parte de las publicaciones se presentan únicamente dos, o una sola columna[8], ya que el nivel de significación se indica mediante asteriscos (dos asteriscos para niveles inferiores al 1 por mil, un asterisco para niveles inferiores al 5% y sin asterisco para niveles superiores al 5%) de la siguiente forma:

	e^b
SEXO(1)	,724**
ESTUDIOS(1)	2,627**
ESTUDIOS(2)	3,419**
Constante	,145**

*Ns=0,05
**Ns=0,001

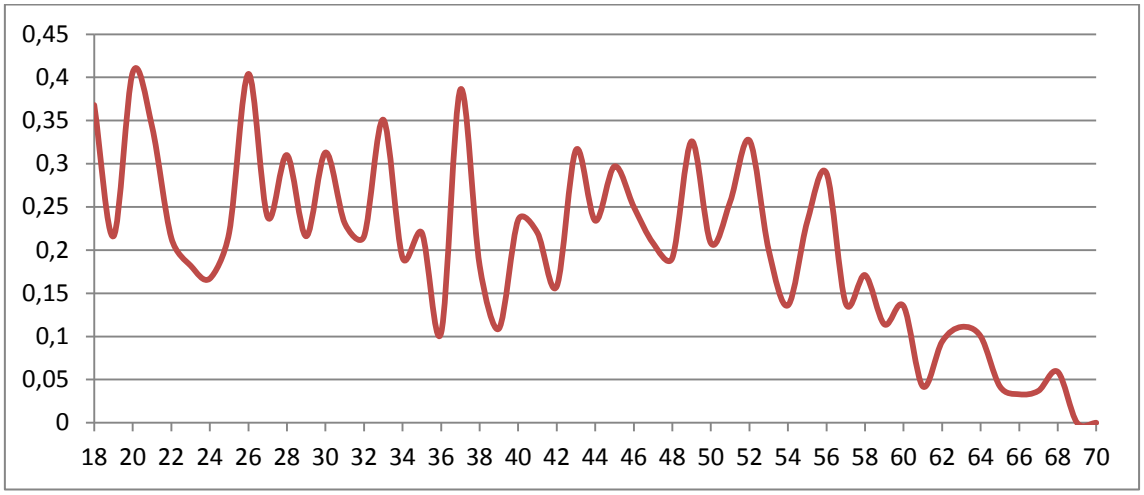
El uso de los modelos de regresión en sociología se centra en la interpretación de los coeficientes Odd Ratio. En este caso observamos que el nivel de estudios tiene una relación importante con la participación en huelgas, quienes tienen estudios universitarios hacen 3,4 veces más huelga que quienes tienen estudios básicos. Ser mujer reduce la participación en una huelga, sólo hay tres mujeres que hacen huelga por cada cuatro hombres que la hacen. ($3/4=0{,}75\approx0{,}72$). No obstante, esta es una lectura de provisional de datos. Como se verá durante este capítulo, los coeficientes se interpretan dentro de un modelo explicativo. Antes de introducir la especificación de modelos vamos a avanzar en la introducción de nuevas variables, pero no cualitativas o de categoría, sino de intervalo como es la edad.

[8] El valor de b es fácilmente deducible de e^b. Es simplemente su logaritmo. Por lo general en investigación social, se interpreta e^b que es lo OR.

6. Variables dependientes de intervalo

En regresión logística además de variables de categoría podemos considerar variables de intervalo. En el caso que estamos estudiando vamos a considerar el efecto de la edad en la participación como huelguista.

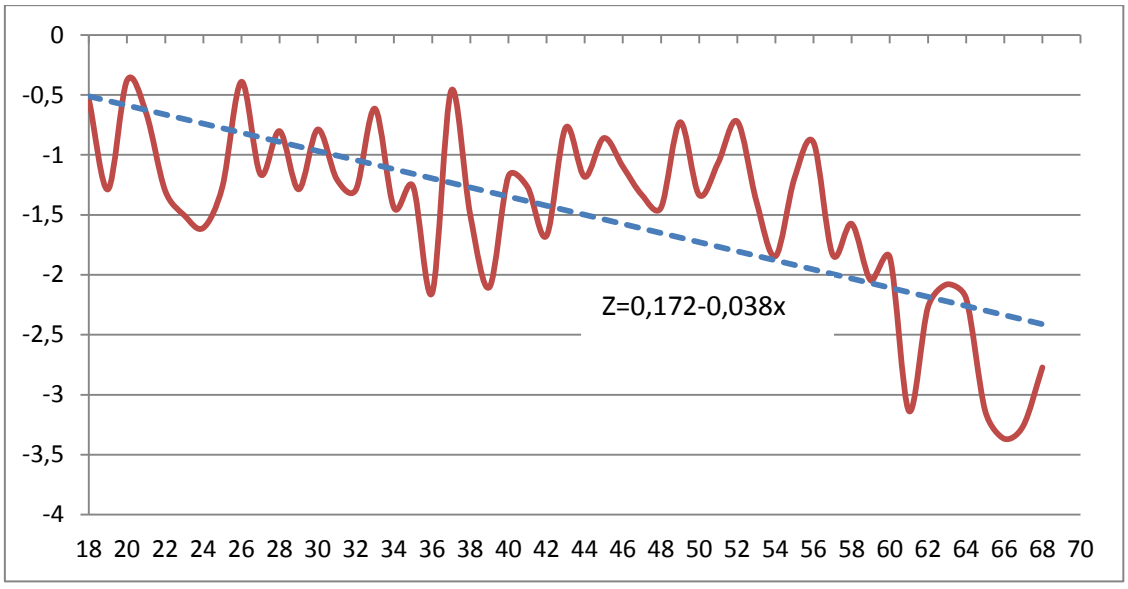
Gráfico 3. Proporción de participantes en huelgas por edad.



El gráfico, realizado con los datos de la encuesta, señala como tendencia general que la participación en las huelgas se reduce con la edad.

Podemos presentar los datos también con el Logit de participación y obtener una función lineal Z=a+bx para determinar el Logit en función de la edad.

Gráfico 4. Logit de participantes por edad



El programa de ordenador nos ofrece el siguiente resultado:

	B	E.T.	Wald	gl	Sig.	Exp(B)
EDAD	-0,038	0,003	120,306	1	0	0,963
Constante	0,172	0,148	1,357	1	0,244	1,188

De forma que podemos determinar el logit mediante la función de edad:

z=0,172-0,038x.

La ecuación obtenida es ligeramente distinta que la recta de regresión lineal que hubiéramos obtenido por el método de mínimos cuadrados. El método de mínimos cuadrados que se utiliza en regresión lineal busca minimizar la suma de las distancias (diferencias al cuadrado) entre los valores observados y estimados. En regresión lineal el procedimiento que utilizamos es diferente y se basa en los métodos de máxima-verosimilitud. Sin necesidad de entrar en teoría inferencial diremos que este método evalúa un conjunto de coeficientes que hayan podido generar los datos observados y selecciona aquel valor del parámetro que tiene la probabilidad más elevada de haber generado la serie de datos observada.

Como podemos observar b es negativo, quiere decir que a mayor edad menor realización de huelgas. El término e^b=0,963, tiene una lectura diferente respecto al caso de variables de categoría. En este caso es la pendiente, o relación entre dos valores. Por cada año de edad la proporción de huelga será un 96,3% respecto del valor anterior. La pendiente es casi 1, quiere decir que la "caída" es lenta.

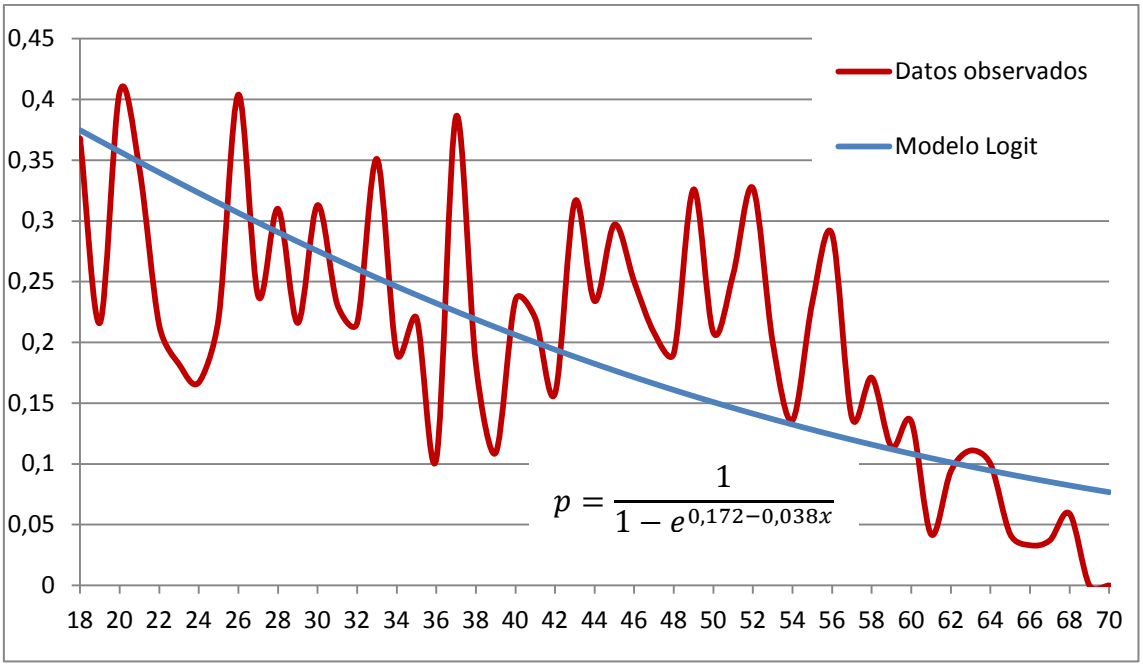
Los coeficientes nos permiten construir la siguiente relación:

z=0,172-0,038x

$$p = \frac{1}{(1 - e^{-z})} = \frac{1}{1 - e^{0,172-0,038x}}$$

Dando valores, podemos contrastar los resultados de nuestro modelo con los datos observados:

Gráfico 5. Proporción de participantes en huelgas por edad. Observado y pronosticado.



La observación de ambas series nos hace preguntarnos por la adecuación del modelo a los datos, cuestión que introduciremos en la siguiente sección. Pero presentaremos ahora dos coeficientes que miden de forma similar a cómo mide el coeficiente de correlación de Pearson la asociación entre la variables independientes y dependientes. Los programas de ordenador ofrecen distintos coeficientes con este propósito, que tratan de emular el célebre coeficiente de determinación de Pearson, y que se denominan de forma genérica: Pseudo-R^2.

En este caso, se ha obtenido el de Cox y Snell con un valor de R^2_{Cox}=0,054. Este coeficiente toma valores entre 0 y 1 de forma que 0 indicaría un efecto muy bajo de las variables independientes, mientras que en la proximidades de 1 mostraría un efecto considerable. Sin embargo, como se explicará más adelante, este coeficiente no puede llegar a valer 1. Por eso se utiliza el R^2 de Nagelkerke, que es el valor del R^2 de Cox y Snell estandarizado sobre el valor máximo que éste podría tomar. De esta forma se garantiza que se pueda interpretar su valor entre 0 y 1. El valor obtenido en este caso,$R^2_{Nagelkerke}$=0,088, señala el "pésimo" ajuste que han tenido nuestros datos.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2262,560[a]	,054	,088

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Junto al valor de los Pseudo R^2 el programa nos ofrece también el estadístico "-2 log de la verosimilitud". Este estadístico se refiere a la probabilidad con la que se determina que el coeficiente, en este caso b, produce los datos. Como hemos señalado la estimación del coeficiente se produce por el método de máxima verosimilitud.

Explicado de forma sencilla el programa evalúa varios valores de "b" (y también de a) para determinar cuál -o qué combinación- es el que tiene la mayor probabilidad de haber producido los datos observados.

Un ejemplo para acercarse a la noción de máxima verosimilitud

Supongamos que en una ciudad hay dos asociaciones culturales. La asociación *Godot*, apoya el teatro y está compuesta por un 45% de mujeres y un 55% de hombres. La asociación S*anti Andía* practica los talleres literarios y está compuesta por un 65% de mujeres y un 35% de hombres. Un sociólogo ha sido invitado a impartir una conferencia, pero como es un poco despistado y ya cuando está en el salón de actos no recuerda en qué asociación se encuentra. Si pregunta por la asociación resultará muy descortés. Decide poner en práctica sus conocimientos de estadística para salir de la situación. Se fija en las dos primeras fila y anota el sexo de los 24 primeros espectadores: 12 son hombres y 13 son mujeres.

El 54% son mujeres y el 46% son hombres. Saca un lápiz y después de garabatear unos números comienza diciendo, entre un fuerte aplauso, "El teatro no ha muerto…"

¿Qué ha hecho nuestro conferenciante?, simplemente calcular la probabilidad que tiene la sala de tener una relación de sexos observada.

Si se encuentra en la asociación *Godot*, la probabilidad de que 24 asistentes 13 sean mujeres es:

$$p_{Godot} = \binom{24}{13} 0{,}45^{13} \cdot 0{,}55^{11} = 0{,}108$$

Si está en la asociación *Santi Andía*

$$p_{Santi\ Andía} = \binom{24}{13} 0{,}65^{13} \cdot 0{,}35^{11} = 0{,}089$$

Es más probable que se encuentre en la asociación *Godot*, y tiene una ventaja de 0,108/0,089=1,21

Nuestro conferenciante ha aplicado el método de máxima verosimilitud. En este caso ha sido sencillo, porque sólo había dos posibilidades.

La probabilidad que tiene un coeficiente o un modelo de haber generado unos datos se denomina verosimilitud, y suele indicarse con "L", que es la primera letra de la palabra Likelihood o verosimilitud en inglés. Las probabilidades son números entre 0 y 1, pero los valores que se manejan como verosimilitud son tan extremadamente pequeños que se trabaja con el logaritmo neperiano (Ln) de dichas cantidades. Como el logaritmo de un número menor que 1 es negativo, se toma el valor -2Ln(L) como indicador de la verosimilitud[9]. Muchos programas y textos utilizan -2LL como acrónimo. (La primera L indica logaritmo (neperiano) y la segunda verosimilitud.

Mayores valores de -2LL muestran una verosimilitud del modelo menor. La verosimilitud de modelo se calcula mediante el ratio entre la verosimilitud del modelo propuesto sobre un modelo que contiene únicamente la constante, es decir, en el que las variables independientes no tienen efecto. Así, la verosimilitud es una medida relativa que nos permite contrastar modelos. Para el contraste de modelos, denominamos *modelo base* a aquel que sólo contiene la constante. El modelo que vamos a contrastar cuando contiene todas las variables de denomina *saturado*.

Vamos a observar la estimación máximo-verosímil para el modelo Logit con la variable edad. En primer lugar, el programa estima la constante "a". Para la estimación los programas estadísticos utilizan algoritmos iterativos que paran cuando los nuevos valores que obtienen no se diferencian de los anteriores. En este caso realiza cuatro cálculos seguidos para el valor de la constante. Cuando la diferencia entre valores es menor de 1/1000 el programa deja de calcular. En este caso estima el valor de a=-1,463, con una medida de verosimilitud:

-2LL=2401,107.

Historial de iteraciones[a,b,c]

Iteración		-2 log de la verosimilitud	Coeficientes Constant
Paso 0	1	2419,432	-1,248
	2	2401,170	-1,450
	3	2401,107	-1,463
	4	2401,107	-1,463

a. En el modelo se incluye una constante.

b. -2 log de la verosimilitud inicial: 2401,107

c. La estimación ha finalizado en el número de iteración 4 porque las estimaciones de los parámetros han cambiado en menos de ,001.

[9] Por ejemplo una probabilidad de 1/1000000 tiene un valor en $-2Ln(1x10^{-6})=27,631$. Y una probabilidad de $1x10^{-12}$ tiene un $-2Ln(1x10^{-12})=55,262$. Observe que en las tablas los valores de -2LL están en torno al número dos mil. Si no fuera por la transformación serían valores inapreciables. (Multiplicar el Ln por 2 tiene el efecto de elevar el valor de L al cuadrado. Recuerde que $2Ln(L)=Ln(L^2)$

Una vez ajustado un modelo sin variables independientes, ajustamos un modelo con variables dependientes, el modelo saturado.

Historial de iteraciones[a,b,c,d]

Iteración		-2 log de la verosimilitud	Coeficientes	
			Constant	P26
Paso 1	1	2315,629	-,302	-,020
	2	2264,477	,049	-,034
	3	2262,564	,166	-,037
	4	2262,560	,172	-,038
	5	2262,560	,172	-,038

a. Método: Introducir

b. En el modelo se incluye una constante.

c. -2 log de la verosimilitud inicial: 2401,107

d. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Después de varias iteraciones el programa ajusta el modelo con los coeficientes:

a=0,172 y b=-0,038.

Este modelo tiene una verosimilitud -2LL=2262,560

La relación de verosimilitud, diferencia entre los logaritmos, sigue una distribución Ji-cuadrado. Los programas nos suelen ofrecer un test, para indicar si el modelo que utilizamos es significativo respecto al modelo base.

La diferencia en verosimilitud es:

2262,50-2401,107=138,547

que es el valor Ji-cuadrado asociado al modelo. En este caso resulta significativo al 99,99%

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	138,547	1	,000
	Bloque	138,547	1	,000
	Modelo	138,547	1	,000

Si volvemos ahora a los resultados, podemos comprobar el significado de los pseudo-R^2:

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2262,560[a]	,054	,088

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Vamos a denominar L_S al valor de la verosimilitud del modelo con todas las variables -modelo saturado- y L_0 a la verosimilitud del modelo que sólo tiene la constante -modelo base-. El efecto de la variable en el modelo vendrá dado por la diferencia en verosimilitud. Como manejamos logaritmos, la diferencia se convierte en ratio[10].

Llamemos L_m al valor de verosimilitud de un modelo que considera determinada/s variable/s. Para nuestros datos, si calculamos dicho ratio entre los logaritmos de la verosimilitud del modelo con la variable edad respecto al modelo sólo con constante, tenemos que:

$$R^2_{McFadden} = \frac{Ln(L_m)}{Ln(L_0)} = \frac{-1131{,}28}{-1200{,}554} = 0{,}9422987$$

El valor de $Ln(L_m)$ y $Ln(L_0)$ lo obtenemos al dividir entre -2, los valores que nos ofrece el programa para -2LL en ambos modelos.

Nuestro modelo, que ajusta la variable edad, es un 94,3% mejor (más verosímil) que el modelo sin variable independiente. Por lo tanto 1-0,9422987=0,057. La variable edad tendría una capacidad de explicar el modelo del 5,7%, que es el valor del Pseudo-R^2 de McFadden[11].

Sin embargo, a diferencia de lo que ocurre con el coeficiente de determinación de Pearson, este coeficiente no es comparable entre distintos estudios. El coeficiente de McFadden sólo nos permite comparar modelos distintos para el mismo conjunto de datos. Por ello nos puede orientar para introducir o eliminar variables en función de su capacidad de ajuste. Sin embargo no podemos interpretarlo como una medida de asociación. Téngase en cuenta que el modelo base, que sirve como denominador, y que permitiría estandarizar los valores, resulta diferente para cada conjunto de datos. Por ello se han elaborado otros coeficientes que intentan evitar este problema.

No resulta apropiado en un texto introductorio el desarrollo de los coeficientes R^2, y simplemente los presentaremos como una extensión del razonamiento de la diferencia

[10] Recuérdese que la diferencia entre A-B=Ln(A)/Ln(B)
[11] Ni el programa SPSS ni PSPP calculan este estadístico.

entre verosimilitud. El coeficiente R^2 de Cox y Snell en notación matemática adquiere la formulación de la media geométrica de la relación entre verosimilitudes[12].

$$R^2_{Cox} = 1 - \left(\frac{L_s}{L_0}\right)^{\frac{2}{n}}$$

El valor de este coeficiente está limitado por $1-(L_o)^{2/n}$

Teniendo esto en cuenta el R^2 de Nagelkake se resuelve:

$$R^2_{Nagelkake} = \frac{1 - \left(\frac{L_s}{L_0}\right)^{\frac{2}{n}}}{1 - (L_o)^{\frac{2}{n}}}$$

El R^2_{Cox} podemos calcularlo con la expresión equivalente a partir de los Logaritmos de verosimilitud

$$R^2_{Cox} = 1 - e^{\frac{(-2Ln(L_s) - 2Ln(L_0))}{n}} = 1 - e^{\frac{(2262{,}56 - 2401{,}107)}{2479}} =$$

$$1 - e^{-0{,}05588826} = 1 - 0{,}9456448 = 0{,}0543552 \approx 0{,}054$$

Y el valor máximo de R^2_{Cox}

$$Max = 1 - e^{2Ln(L_o)} = 1 - e^{-2401{,}107} = 1 - 0{,}380363 = 0{,}619637$$

Luego

$$R^2_{Nagelkake} = \frac{0{,}0543552}{0{,}619637} = 0{,}087772 \approx 0{,}088$$

No obstante, conviene tener presente que, si bien en teoría estos coeficientes pueden alcanzar valores elevados, resulta difícil en los ajustes que incluso puedan superar valores del 0,3 (30%). No son estrictamente medidas de asociación sino de probabilidad que nos orientan sobre la ganancia que nos produce el modelo sobre situaciones en las que no tenemos modelos. Más adelante veremos otras formas de evaluar el ajuste del modelo a los datos.

Estos indicadores nos muestran qué ganamos utilizando un modelo, pero no nos indican qué capacidad tenemos de predecir los datos, algo que si ofrecía el coeficiente de determinación.

El coeficiente de Cox y Snell, puede expresarse también en términos de Ji-cuadrado como:

$$R^2_{Cox} = \frac{\chi^2}{\chi^2 + n} = \frac{138{,}547}{138{,}547 + 2484} = 0{,}053$$

[12] La función de verosimilitud se construye mediante el producto de n (número de casos) términos.

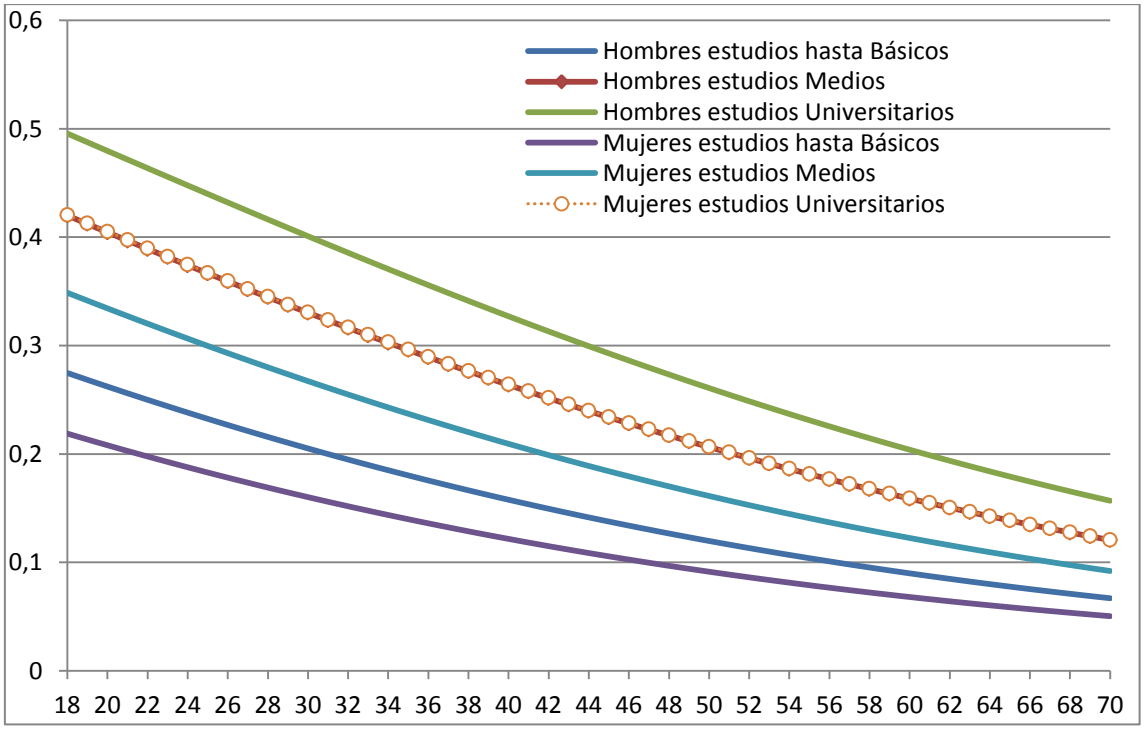
7. La generalización del modelo a varias variables

Como podemos observar, si bien el modelo refleja la tendencia de la serie y el efecto de la variable independiente, sin embargo los resultados particulares en algunos casos quedan muy apartados respecto a los valores pronosticados. Vamos a observar ahora el comportamiento conjunto con tres variables independientes de distinta naturaleza. En concreto con sexo, estudios y edad. El resultado que obtenemos es:

	B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
							Inferior	Superior
Edad	-,032	,004	75,361	1	,000	,969	,962	,976
Sexo(1)	-,302	,108	7,884	1	,005	,739	,599	,913
Estudios			50,617	2	,000			
Estudios(1)	,648	,130	24,943	1	,000	1,911	1,482	2,464
Estudios(2)	,953	,138	47,558	1	,000	2,594	1,978	3,401
Constante	-,395	,193	4,187	1	,041	,674		

En este caso se ha añadido también la estimación del intervalo para los coeficientes. Para una interpretación del modelo, podemos representar las ecuaciones para cada combinación de categorías. (Nótese que en el gráfico se superponen los valores de hombres con estudios medios y de mujeres con estudios universitarios).

Gráfico 6. Representación de ecuaciones para combinaciones de categorías



Ahora vamos a mostrar la utilidad de los Odds para expresar las relaciones que se establecen entre variables de categoría en términos algebraicos.

Los coeficientes muestran que la participación en huelgas está muy relacionada con el nivel de estudios, es la variable que mayor capacidad de determinación tiene: quienes tienen estudios universitarios hacen 3 veces más huelga que quienes tienen menores niveles de estudios. Los hombres hacen más huelga que las mujeres: e^b=0,739 puede expresarse de forma aproximada señalando que sólo hay tres mujeres que hacen huelga por cada cuatro hombres (3/4=0,75). La edad reduce la participación en huelgas, pero lo hace de forma muy lenta: el coeficiente 0,969 quiere decir que por cada año de edad se reduce un 3% la proporción de huelguistas.

Todos los coeficientes son significativos. Sin embargo, como podemos ver el ajuste del modelo resulta muy discreto. Los valores de Pseudo R^2 han mejorado, con la introducción de nuevas variables. Sin embargo el problema reside en la adecuación teórica del modelo. Hemos relacionado un conjunto de variables sin ninguna suposición sobre el efecto causal que puedan tener.

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2196,148[a]	,076	,123

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

7.1. Introducción de variables por pasos

En el apartado anterior hemos introducido las tres variables a la vez. Por regla general los programas de ordenador permiten sistemas de introducción de variables de forma progresiva (stepwise). Estos sistemas deciden que variables se introducen en cada paso y cuales finalmente quedan fuera del modelo. Como se verá en la parte final de este texto no se aconseja el uso de procedimientos automáticos para la selección de variables en los modelos. Se recomienda introducir o sacar variables siempre que exista fundamento teórico para ello.

En el caso que estamos analizando de tres variables, hemos solicitado la introducción de las variables paso a paso. Los resultados finales han sido los mismos, las tres variables han sido introducidas y los coeficientes idénticos a los obtenido en la introducción de todo el conjunto de variables.

La tabla siguiente resume el proceso:

Modelo si se elimina el término[a]

Variable		Log verosimilitud del modelo	Cambio en -2 log de la verosimilitud	gl	Sig. del cambio
Paso 1	Edad	-1199,397	143,613	1	,000
Paso 2	Edad	-1144,908	85,741	1	,000
	Estudios	-1127,915	51,754	2	,000
Paso 3	Sexo	-1102,045	7,942	1	,005
	Edad	-1140,181	84,215	1	,000
	Estudios	-1124,451	52,754	2	,000

a. Según las estimaciones condicionales de los parámetros

Se ha seleccionado en primer lugar la variable edad, en segundo lugar se ha introducido la variable estudios y la última en entrar ha sido la variable sexo. La introducción de cada variable ha ido mejorando el ajuste:

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2255,180[a]	,054	,087
2	2204,076[a]	,073	,119
3	2196,148[a]	,076	,123

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

¿Cuánto ha mejorado el modelo? Es lo que nos indica el estadístico -2Ln(v), y que, como hemos visto, podemos interpretar mediante los pseudo-R^2

7.2 La capacidad predictiva del modelo

En apartados anteriores hemos visto la información que proporcionan las funciones de verosimilitud para orientarnos en la incorporación de nuevas variables para contrastar el grado de significación del modelo. Aunque el uso no es frecuente en ciencias sociales, en algunas especialidades, especialmente en el campo de la salud, se pide una capacidad predictiva a los modelos Logit. Ahora veremos una herramienta que ayuda a valorar la capacidad predictiva del modelo respecto a los datos: la *Tabla de Clasificación*.

La Tabla de Clasificación es una tabla de doble entrada que cruza los valores observados de la variable con los valores predichos por el modelo considerado para cada uno de los casos. Para el modelo sexo+edad+estudios que estamos analizando, el resultado es el siguiente:

Tabla de clasificación[a]

			Pronosticado		
			Huelga		Porcentaje
	Observado		No	Sí	correcto
Paso 1	Huelga	No	2014	0	100,0
		Sí	465	0	,0
	Porcentaje global				81,2

a. El valor de corte es ,500

Por una parte, la tabla nos muestra que en los datos observados[13] hay 2014 entrevistados que han declarado no haber realizado huelga y 465 que sí. Por otra parte, la tabla nos dice cuál es el pronóstico para los diferentes casos respecto a la característica que estamos estudiando.

¿Cómo hacemos el pronóstico? Para cada uno de los casos de la encuesta calculamos la probabilidad que tiene en función de sus características. Por ejemplo, observemos los 10 primeros entrevistados, según las variables que estamos considerando:

Número de Cuestionario	Nivel de Estudios	Sexo	Edad	Huelga
1566	Hasta Básicos	Hombre	45	Sí
1567	Medios	Hombre	25	No
1568	Hasta Básicos	Mujer	48	No
1569	Hasta Básicos	Mujer	87	No
1570	Medios	Mujer	18	No
1571	Hasta Básicos	Hombre	22	No
1572	Universitarios	Hombre	37	No
1573	Medios	Mujer	33	No
1574	Hasta Básicos	Hombre	78	No
1575	Hasta Básicos	Hombre	27	No

Por ejemplo, el primer entrevistado -cuestionario 1566- es un hombre de 45 años con estudios básicos o inferiores (sabemos -valor observado- que ha hecho huelga). Con estos datos podemos calcular la probabilidad que tiene de hacer huelga, a partir de los coeficientes "b" calculados:

$Z=a+b_1edad+b_2sexo+b_3estudios = Z=-0{,}395 - 0{,}032(45) - 0{,}302(0) + 1(0)$

Z=-1,819

Y la probabilidad que calculamos será:

[13]En este análisis los 2484 casos se han quedado en 2479, al haberse eliminado del análisis unos pocos casos en los que no se podía determinar el nivel de estudios.

$$p = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(-1{,}819)}} = 0{,}139$$

Es decir, al entrevistado 1566 le asignamos una probabilidad de 13,9% de hacer huelga. Como esta probabilidad es menor del 50% pronosticamos que no habrá hecho huelga.

Procediendo de la misma forma para los 10 primeros entrevistados tenemos que:

Número de Cuestionario	Nivel de Estudios	Sexo	Edad	Huelga	Probabilidad Pronosticada
1566	Hasta Básicos	Hombre	45	Sí	0,13954
1567	Medios	Hombre	25	No	0,36854
1568	Hasta Básicos	Mujer	48	No	0,09829
1569	Hasta Básicos	Mujer	87	No	0,03075
1570	Medios	Mujer	18	No	0,34995
1571	Hasta Básicos	Hombre	22	No	0,2514
1572	Universitarios	Hombre	37	No	0,35145
1573	Medios	Mujer	33	No	0,25087
1574	Hasta Básicos	Hombre	78	No	0,05399
1575	Hasta Básicos	Hombre	27	No	0,2228

De estos diez primeros casos, ninguno obtiene probabilidad superior al 0,5, por lo tanto todos son considerados como casos sin la característica. Si observamos la tabla de clasificación, podemos ver que ninguno de los cuestionarios ha sido pronosticado como huelguista.

Para el modelo concreto de tres variables que hemos expuesto, el caso más favorable sería un hombre joven de 18 años con estudios superiores. Para este entrevistado la probabilidad asignada al caso más favorable sería:

$$p = \frac{1}{1 + e^{-(-0{,}395 - 0{,}032(18) - 0{,}302(0) + 953(1))}} = 0{,}496$$

Esa es la probabilidad más alta que este modelo asigna a un individuo[14], por eso, como la tabla de clasificación utiliza como punto de corte p=0,5 no pronostica ningún caso como huelguista. Algunos programas ofrecen el siguiente gráfico, que permite observar la probabilidad que el modelo asigna y el valor observado de los casos:

[14] En la práctica ni siquiera existe el caso de joven de 18 años con estudios universitarios terminados. Se trata del hipotético caso con mayor probabilidad.

```
             Observed Groups and Predicted Probabilities

     160 +                                                                                                    +
         I     N                                                                                              I
         I     N                                                                                              I
F        I     N                                                                                              I
R    120 +     NS                                                                                             +
E        I    NNN                                                                                             I
C        I    NNN                                                                                             I
U        I    NNNNSS                                                                                          I
E     80 +    NNNNSN   SS  SS       S                                                                         +
N        I   NNNNNNNSSSNNS SSSSS S SS                                                                         I
C        I   NNNNNNNNNNNNSSNNNSSSN NSS   S                                                                    I
Y        I   NNNNNNNNNNNNNSNNNNNSNSNNN SNS                                                                    I
      40 +   NNNNNNNNNNNNNNNNNNNNNSNNNSNNS SS S                                                               +
         I   NNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNSN S  S                                                          I
         I   NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNSSNS                                                         I
         I   NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNS                                                       I
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        ,1        ,2        ,3        ,4        ,5        ,6        ,7        ,8        ,9         1
  Group:  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS

          Predicted Probability is of Membership for Sí
          The Cut Value is ,50
          Symbols: N - No
                   S - Sí
          Each Symbol Represents 10 Cases.
```

A partir de la tabla de clasificación, que relaciona los casos observados con los pronosticados, podemos obtener distintas medidas de la capacidad predictiva del modelo. Vamos a llamar A, B, C y D a cada una de las cuatro casillas que componen la tabla de clasificación (N representa el total de casos):

		Pronosticado		
		No	Si	
Observado	No	A	B	
	Si	C	D	
				N

La primera medida que podemos establecer es el porcentaje de casos que el modelo predice perfectamente, aquéllos que pronostica como negativos y son efectivamente negativos (A) y aquellos que pronostica como positivos y son efectivamente positivos (D). El porcentaje correcto para nuestros datos será:

$$\frac{(A+D)}{N}x100 = \frac{2014+0}{2479}x100 = 81{,}2\%$$

Como vemos, en este caso el porcentaje de clasificación correcta es elevado. Sin embargo nuestro modelo resulta poco útil ya que ha otorgado a todos los elementos el mismo valor. Por ello se utilizan otras medidas que contemplan los falsos y los verdaderos positivos y negativos.

Falsos positivos, son los casos que el modelo clasifica con la característica pero que no la tienen:

$$FalsosP = \frac{B}{(A+B)}x100 = \frac{0}{(2014+0)}x100 = 0\%$$

Falsos negativos, casos que teniendo la característica el modelo no reconoce como tales:

$$FalsosN = \frac{C}{(C+D)}x100 = \frac{465}{(465+0)}x100 = 100\%$$

Verdaderos positivos, casos clasificados con la presencia de la característica cuando efectivamente estos la tienen. (Para nuestro ejemplo no puede calcularse, porque no hay ningún positivo pronosticado).

$$VerdaderosP = \frac{D}{(B+D)}x100 = \frac{0}{(0+0)}x100$$

Verdaderos negativos, casos clasificados en la categoría de ausencia de característica, cuando efectivamente no la tienen.

$$VerdaderosN = \frac{A}{(A+C)}x100 = \frac{2014}{(2014+465)}x100 = 81{,}2\%$$

Para nuestros datos, si se atiende a falsos positivos el modelo funciona, y si consideramos la medida de falsos negativos el modelo no funciona. Dada la ambigüedad de este tipo de medidas se han definido la *sensibilidad* y la *especificidad* del modelo.

La *sensibilidad del modelo* se refiere a la capacidad que tiene éste para detectar como positivos los casos que poseen la característica. En términos coloquiales, si al modelo le presentamos sólo casos positivos, la sensibilidad determina la capacidad que tiene el modelo de no equivocarse. La sensibilidad queda definida como:

$$Sensibilidad = \frac{VerdaderosP}{VerdaderosP + FalsosN}$$

Y a través de las frecuencias de las casillas puede indicarse de forma operativa como:

$$\frac{D}{(C+D)}x100 = \frac{0}{(465+0)}x100 = 0\%$$

La *especificidad del modelo* se refiere a la capacidad que tiene éste para discriminar correctamente los casos que no poseen la característica. Es decir, sobre un conjunto de casos que no poseen la característica, determina en qué grado no va a confundirlos con casos que poseen la característica. La especificidad la definimos como:

$$Especificidad = \frac{VerdaderosN}{VerdaderosN + FalsosP}$$

Que podemos indicar, también, como:

$$\frac{A}{(A+B)}x100 = \frac{2014}{(2014+0)}x100 = 100\%$$

Un modelo con buena capacidad predictiva debería tener valores altos tanto de sensibilidad como de especificidad.

El modelo propuesto tiene una sensibilidad nula, es incapaz de detectar los casos que contienen la característica: se trata de un modelo malo. Como veremos en los siguientes apartados, en ciencias sociales la validez de un modelo estadístico viene determinada por su validez teórica antes que por su validez como predictor. El modelo que hemos utilizado hasta ahora no ha sido elaborado con el mínimo soporte teórico, y eso se nota también en su ajuste empírico.

III. ELABORACIÓN Y CONTRASTE DE UN MODELO

En la elaboración de un modelo logit debemos tratar de introducir en un primer momento todas las variables que se consideran a nivel teórico como relevantes. En nuestro caso, además de las variables sociodemográficas -sexo y edad- hemos considerado el nivel de estudios, que como hemos visto tiene un efecto importante. Sin embargo no hemos introducido otras variables que resultan cruciales, como son la relación con la actividad y la ideología política.

En primer lugar debemos especificar cuál es el dominio de nuestro modelo. La realización de una huelga implica tener la condición de trabajador (40,7%) o una vinculación con la actividad (parados que han trabajado 22,4%). También podemos incluir al grupo de estudiantes (4,6%).

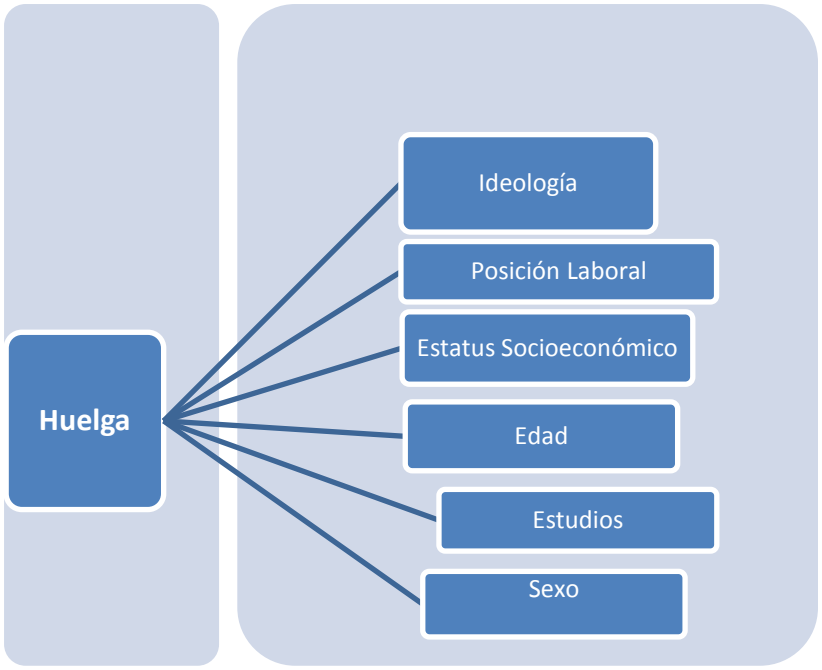
P33

	Frecuencia	Porcentaje
Trabaja	1011	40,7
Jubilado/a o pensionista (anteriormente ha trabajado)	459	18,5
Pensionista (anteriormente no ha trabajado)	79	3,2
Parado/a y ha trabajado antes	556	22,4
Parado/a y busca su primer empleo	31	1,2
Estudiante	115	4,6
Trabajo doméstico no remunerado	222	8,9
Otra situación	9	,4
N.C.	2	,1
Total	2484	100,0

A partir de aquí, vamos a restringir el conjunto de datos al grupo compuesto por trabajadores, parados y estudiantes. Es decir las tablas que siguen no se refieren a toda la población española sino únicamente al colectivo de referencia (67,7%) que constituye el universo de posibles huelguistas.

*L*a participación en una huelga es una acción en la que intervienen distintos factores, pero que podemos resumir en ideología política y posición laboral. La huelga es un derecho laboral pero que combina también la protesta política, como es en el caso de una huelga general. Como instrumento de protesta está vinculado a la ideología política, tradicionalmente la izquierda apoya de forma más explícita la huelga que la derecha. Como derecho laboral, la huelga se asocia a la posición laboral. Tienen mayor posibilidad de realizar huelga quienes están en posiciones más estables que quienes se encuentran en posiciones de mayor precariedad en la medida en que una huelga puede tener mayores repercusiones laborales y salariales.

De forma sintética podemos representar el modelo:



Una vez determinadas las variables, en primer lugar las examinamos detalladamente para hacerlas operativas. Como veremos, tenemos que observar si presentan una distribución extraña, con valores "outliers", en el caso de variables de intervalo, o con categorías con un número de efectivos muy reducido en el caso de variables cualitativas. También deberemos tomar decisiones respecto a categorías residuales y de no respuesta.

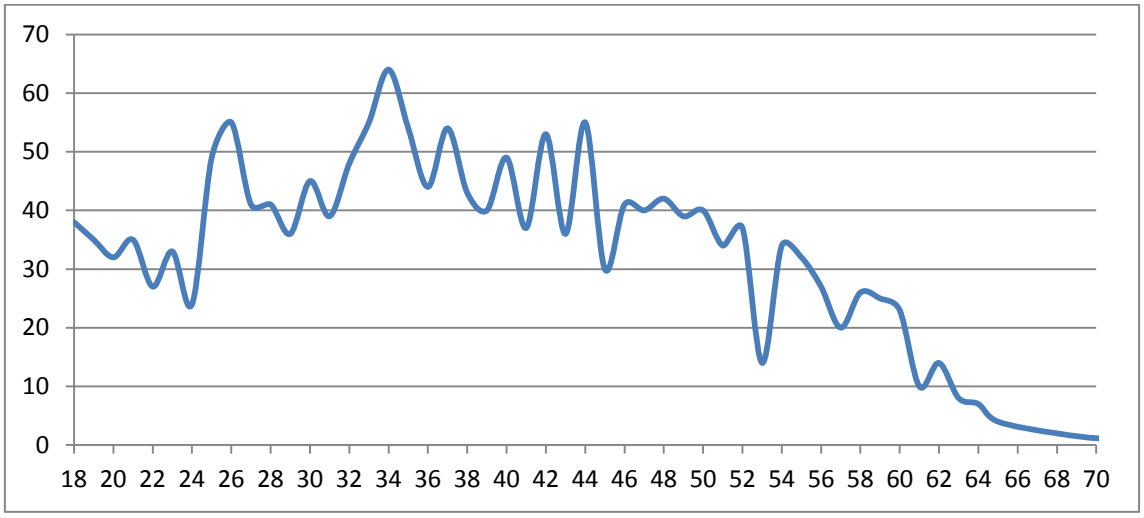
Las variables que vamos a analizar son:

Sexo, variable p25 en el cuestionario.

P25

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Hombre	1221	49,2	49,2	49,2
	Mujer	1263	50,8	50,8	100,0
	Total	2484	100,0	100,0	

Edad, pregunta 26 del cuestionario.



Nivel de estudios. Esta variable viene generada por el propio CIS, como combinación de las preguntas 27 y 28. El resultado de la variable ESTUDIOS es:

Estudios del entrevistado

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Sin estudios	27	1,6	1,6	1,6
Primaria	641	37,4	37,4	39,0
Secundaria	277	16,2	16,2	55,2
F.P.	337	19,7	19,7	74,8
Medios universitarios	168	9,8	9,8	84,6
Superiores	258	15,1	15,1	99,7
N.C.	5	,3	,3	100,0
Total	1713	100,0	100,0	

Para nuestro modelo hemos simplificado la variable en tres categorías, que permiten observar el efecto del nivel de estudios de forma más clara. La agrupación, una vez excluidos los 5 casos que no contestan, ha sido:

		Nivel de Estudios		
		Hasta Básicos	Medios	Universitarios
Estudios del entrevistado	Sin estudios	27		
	Primaria	641		
	Secundaria		277	
	F.P.		337	
	Medios universitarios			168
	Superiores			258
	N.C.			

La variable *condición socioeconómica* nos permite posicionar a los entrevistados según tipos de ocupación:

Condición socioeconómica

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Directores y profesionales	148	8,6	8,6	8,6
	Técnicos y cuadro medios	167	9,7	9,7	18,4
	Pequeños empresarios	76	4,4	4,4	22,8
	Agricultores	30	1,8	1,8	24,6
	Empleados de oficinas y servicios	142	8,3	8,3	32,9
	Obreros cualificados	129	7,5	7,5	40,4
	Obreros no cualificados	185	10,8	10,8	51,2
	Parados	587	34,3	34,3	85,5
	Estudiantes	115	6,7	6,7	92,2
	No clasificables	134	7,8	7,8	100,0
	Total	1713	100,0	100,0	

La variable *estatus socioeconómico*. Es una variable que elabora el propio CIS y que busca reflejar la posición del hogar del entrevistado en estratos sociales. Sus categorías son:

Estatus socioeconómico

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Clase alta/ media-alta	336	19,6	19,6	19,6
Nuevas clases medias	365	21,3	21,3	40,9
Viejas clases medias	243	14,2	14,2	55,1
Obreros cualificados	460	26,9	26,9	82,0
Obreros no cualificados	211	12,3	12,3	94,3
No consta	98	5,7	5,7	100,0
Total	1713	100,0	100,0	

Para facilitar la interpretación de los resultados se eliminarán los 98 casos en los que no ha sido posible realizar dicha clasificación.

La variable *ideología* se obtiene a partir de la pregunta 22, en la que se utiliza el autoposicionamiento en una escala 1(Izquierda) a 10 (Derecha).

P22

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Izquierda	67	3,9	3,9	3,9
	2	91	5,3	5,3	9,2
	3	280	16,3	16,3	25,6
	4	233	13,6	13,6	39,2
	5	392	22,9	22,9	62,1
	6	160	9,3	9,3	71,4
	7	89	5,2	5,2	76,6
	8	69	4,0	4,0	80,6
	9	12	,7	,7	81,3
	Derecha	16	,9	,9	82,3
	N.S.	162	9,5	9,5	91,7
	N.C.	142	8,3	8,3	100,0
	Total	1713	100,0	100,0	

En este caso se han eliminado un volumen importante de efectivos -los 162 que NS más los 142 que NC -el 17,8% de la muestra. Estas respuestas no pueden atribuirse a ningún valor, y por otra parte, esta es una variable fundamental en la explicación de la participación en huelgas. Una solución alternativa podría ser el categorizar la variable ideología en varios grupos -por ejemplo, izquierda, centro y derecha- y añadir como uno o dos grupos más los NS y NC.

Una vez preparadas las variables se examina el comportamiento de las mismas respecto a la variable dependiente. La lectura nos orienta sobre la lectura de los resultados del modelo. Podemos observar qué variables producen grandes diferencias, por ejemplo estudios y escala ideológica, y cuales introducen poca variabilidad, por ejemplo sexo. Es muy importante valorar en esta fase si hay alguna variable que concentre toda la variabilidad en una categoría y si esta categoría tiene un tamaño reducido. En los datos que estamos observando no se aprecia esta situación.

Examinando las variables:

		huelga		huelga	
		No	Sí	No	Sí
		Recuento	Recuento	% de la fila	% de la fila
Sexo	Hombre	673	248	73,1%	26,9%
	Mujer	605	187	76,4%	23,6%
Nivel de Estudios	Hasta Básicos	545	123	81,6%	18,4%
	Medios	441	173	71,8%	28,2%
	Universitarios	289	137	67,8%	**32,2%**
Condición Socioeconómica	Directores y profesionales	131	17	88,5%	11,5%
	Técnicos y cuadro medios	101	66	60,5%	**39,5%**
	Pequeños empresarios	55	21	72,4%	27,6%
	Agricultores	21	9	70,0%	30,0%
	Empleados de oficinas y servicios	99	43	69,7%	30,3%
	Obreros cualificados	82	47	63,6%	**36,4%**
	Obreros no cualificados	128	57	69,2%	30,8%
	Jubilados y pensionistas	0	0	,0%	,0%
	Parados	470	117	80,1%	19,9%
	Estudiantes	79	36	68,7%	31,3%
	Trabajo doméstico no remunerado	0	0	,0%	,0%
	No clasificables	112	22	83,6%	16,4%
Estatus	Clase alta/ media-alta	225	111	67,0%	**33,0%**
	Nuevas clases medias	269	96	73,7%	26,3%
	Viejas clases medias	201	42	82,7%	17,3%
	Obreros cualificados	348	112	75,7%	24,3%
	Obreros no cualificados	153	58	72,5%	27,5%
Escala Ideológica	1,00	28	39	41,8%	**58,2%**
	2,00	41	50	45,1%	**54,9%**
	3,00	164	116	58,6%	41,4%
	4,00	166	67	71,2%	28,8%
	5,00	315	77	80,4%	19,6%
	6,00	142	18	88,8%	11,3%
	7,00	81	8	91,0%	9,0%
	8,00	63	6	91,3%	8,7%
	9,00	9	3	75,0%	25,0%
	10,00	13	3	81,3%	18,8%

	huelga								
	No			Sí			Total		
	Recuento	Media	Desviación típica	Recuento	Media	Desviación típica	Recuento	Media	Desviación típica
P26	1278	39	12	435	38	12	1713	39	12
ideo	1278	4,87	1,75	435	3,64	1,70	1713	4,53	1,82

Seleccionadas, preparadas y estudiadas las variables, realizamos los cálculos con lo programas informáticos a tal efecto. Para dicho análisis tenemos que tener en cuenta dos cuestiones:

- Desde este texto se recomienda introducir todas las variables. Quiere esto decir que no se deja en manos del programa la selección por criterios estadísticos de las mismas. Cuando leamos los resultados ya valoraremos, en función de un criterio teórico, que variables continúan en siguientes fases. Es decir, la decisión es del analista.
- Otra cuestión importante es señalar cuál es la categoría de referencia. Muchos programas por defecto utilizan la última como categoría de referencia. Sin embargo en muchos casos esta categoría puede ser un cajón de sastre "otros", "resto" y su uso como categoría de referencia resulta complejo para la interpretación de resultados. Cuando los programas no permiten seleccionar la categoría de referencia, habrá que cambiar el orden de la codificación.

A continuación comienza el análisis de los resultados. De las diferentes salidas de ordenador que producen los programas vamos a dejar únicamente las que tienen interés. Debajo de cada tabla introducimos un cuadro con algunos comentarios útiles:

El modelo con todas las variables:

Regresión logística

Resumen del procesamiento de los casos

Casos no ponderados[a]		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	1326	77,4
	Casos perdidos	387	22,6
	Total	1713	100,0
Casos no seleccionados		0	,0
Total		1713	100,0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

El programa nos informa del número de casos empleado. Recuérdese que no se emplea toda la muestra sino únicamente se utilizan los cuestionarios del grupo de ocupados, parados y estudiantes (n=1713). En algunas de las variables, se han definido valores perdidos. En total se han empleado 1326 casos para los que se disponía de información de todas las seis variables.

Codificación de la variable dependiente

Valor original	Valor interno
No	0
Sí	1

Muestra la codificación de la variable dependiente. La asistencia a huelga se ha considerado como la característica (Si=1) a investigar.

Codificaciones de variables categóricas

			Codificación de parámetros								
		Frecuencia	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Condición socioeconómica	**Directores y profesionales**	133	,000	,000	,000	,000	,000	,000	,000	,000	,000
	Técnicos y cuadro medios	151	1,000	,000	,000	,000	,000	,000	,000	,000	,000
	Pequeños empresarios	67	,000	**1,000**	,000	,000	,000	,000	,000	,000	,000
	Agricultores	25	,000	,000	**1,000**	,000	,000	,000	,000	,000	,000
	Empleados de oficinas y servicios	118	,000	,000	,000	1,000	,000	,000	,000	,000	,000
	Obreros cualificados	106	,000	,000	,000	,000	1,000	,000	,000	,000	,000
	Obreros no cualificados	146	,000	,000	,000	,000	,000	**1,000**	,000	,000	,000
	Parados	454	,000	,000	,000	,000	,000	,000	1,000	,000	,000
	Estudiantes	92	,000	,000	,000	,000	,000	,000	,000	1,000	,000
	No clasificables	34	,000	,000	,000	,000	,000	,000	,000	,000	1,000
estatus2	**Clase alta/ media-alta**	300	,000	,000	,000	,000					
	Nuevas clases medias	305	1,000	,000	,000	,000					
	Viejas clases medias	204	,000	**1,000**	,000	,000					
	Obreros cualificados	365	,000	,000	1,000	,000					
	Obreros no cualificados	152	,000	,000	,000	1,000					
Nivel de Estudios	**Hasta Básicos**	508	,000	,000							
	Medios	469	1,000	,000							
	Universitarios	349	,000	1,000							
P25	Hombre	729	,000								
	Mujer	597	1,000								

Esta tabla señala la codificación dummy de las variables dependientes. Nótese que la categoría de referencia es aquélla que tiene todos los valores (0)

Bloque 0: Bloque inicial

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-,946	,061	238,872	1	,000	,388

El programa ajusta en primer lugar el modelo base. Aquél que sólo contiene la constante. En este caso a=-0,946

Bloque 1: Método = Introducir

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	206,365	18	,000
	Bloque	206,365	18	,000
	Modelo	206,365	18	,000

A continuación ajusta el modelo con todas las variables. Recordemos que la diferencia entre la verosimilitud del modelo ajustado respecto al modelo base tiene una distribución Ji-cuadrado. En este caso, el valor de Ji-cuadrado nos informa de que el modelo ajustado se diferencia significativamente del modelo base.

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1365,618[a]	,144	,208

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Estos son los coeficientes pseudo-R^2 que nos ofrece este programa.

Tabla de clasificación[a]

			Pronosticado		
			huelga		
	Observado		No	Sí	Porcentaje correcto
Paso 1	huelga	No	899	56	94,1
		Sí	260	111	29,9
	Porcentaje global				76,2

a. El valor de corte es ,500

Con los datos de la tabla de clasificación podemos indicar que:

Sensibilidad=29,9%.

Especificidad=94,1%

	B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
							Inferior	Superior
SEXO(1)	-0,14	0,144	0,952	1	0,329	0,869	0,656	1,152
EDAD	0	0,006	0,001	1	**0,971**	1	0,987	1,013
IDEOLOGÏA	-0,421	0,043	97,454	1	0	**0,657**	0,604	0,714
estatus2			7,898	4	0,095			
estatus2(1)	-0,203	0,263	0,596	1	0,44	0,816	0,487	1,367
estatus2(2)	-0,69	0,325	4,5	1	**0,034**	0,502	0,265	0,949
estatus2(3)	-0,317	0,276	1,325	1	0,25	0,728	0,424	1,25
estatus2(4)	0,133	0,313	0,181	1	0,671	1,142	0,619	2,107
estu			16,585	2	0			
estu(1)	0,523	0,168	9,69	1	0,002	**1,687**	1,214	2,346
estu(2)	0,819	0,212	14,962	1	0	**2,267**	1,498	3,433
CONDICION			38,886	9	0			
CONDICION(1)	1,117	0,347	10,37	1	**0,001**	3,056	1,548	6,031
CONDICION(2)	1,626	0,444	13,444	1	**0**	**5,085**	2,132	12,131
CONDICION(3)	1,447	0,563	6,611	1	**0,01**	**4,248**	1,41	12,797
CONDICION(4)	1,223	0,398	9,427	1	**0,002**	3,396	1,556	7,411
CONDICION(5)	1,854	0,418	19,654	1	**0**	**6,385**	2,813	14,493
CONDICION(6)	1,219	0,381	10,247	1	**0,001**	3,385	1,604	7,142
CONDICION(7)	0,756	0,335	5,109	1	**0,024**	2,131	1,106	4,105
CONDICION(8)	1,294	0,403	10,294	1	**0,001**	3,649	1,655	8,045
CONDICION(9)	0,358	0,554	0,417	1	0,518	1,43	0,483	4,237
Constante	-0,338	0,491	0,474	1	0,491	0,713		

Finalmente obtenemos la tabla con los coeficientes. De dicha tabla extraemos la tabla que presentaremos en los textos:

Modelo con todas las variables

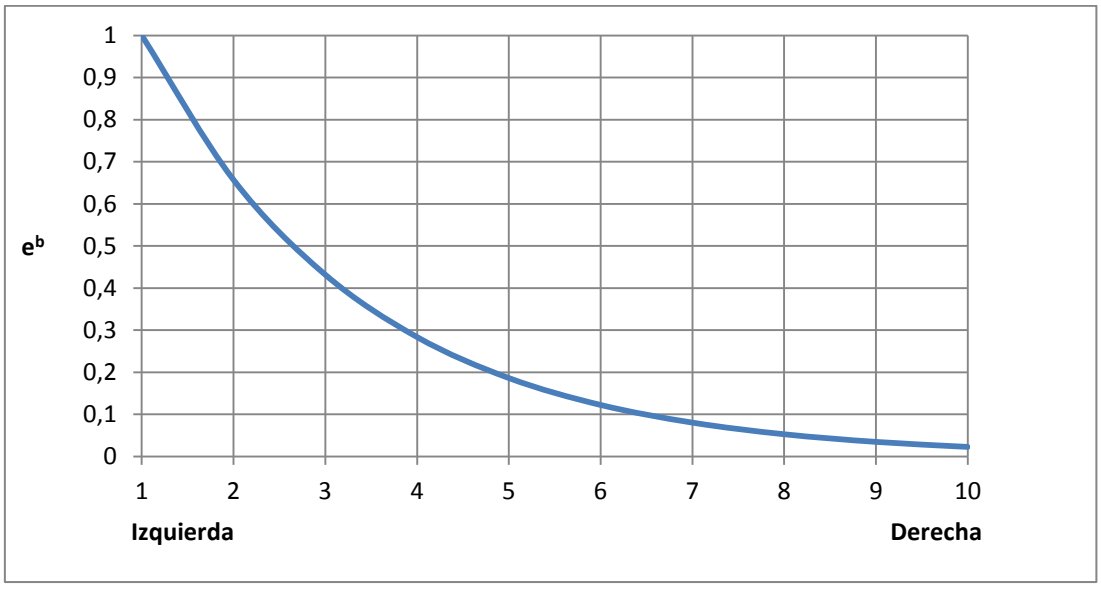
	b	e^b
SEXO (Hombre=0)		
Mujer(1)	-0,14	0,869
EDAD	0	1
IDEOLOGÍA	-0,421	0,657**
ESTATUS (Clase alta/ media-alta=0)		
Nuevas clases medias (1)	-0,203	0,816
Viejas clases medias (2)	-0,69	0,502*
Obreros cualificados (3)	-0,317	0,728
Obreros no cualificados (4)	0,133	1,142
ESTUDIOS (Hasta Básicos=0)		
Medios (1)	0,523	1,687**
Universitarios (2)	0,819	2,267**
CONDICIÓN SOCIOECONÓMICA (Directores y Profesionales=0)		
Técnicos y cuadro medios (1)	1,117	3,056**
Pequeños empresarios (2)	1,626	5,085**
Agricultores (3)	1,447	4,248**
Empleados de oficinas y servicios (4)	1,223	3,396**
Obreros cualificados (5)	1,854	6,385**
Obreros no cualificados (6)	1,219	3,385**
Parados (7)	0,756	2,131*
Estudiantes (8)	1,294	3,649**
No clasificables (9)	0,358	1,43
Constante	-0,338	0,713

Los resultados del modelo indican que las variables sociodemográficas *sexo* o *edad* no tienen influencia en la realización de huelgas: como podemos ver el término e^b puede contener el valor 1 que indica que no hay diferencias. El nivel de significación de ambas variables es claramente superior al 5%. Esta conclusión puede resultar extraña si se tiene en cuenta que lo observado en la primera parte de este texto. Sin embargo, resulta totalmente coherente.

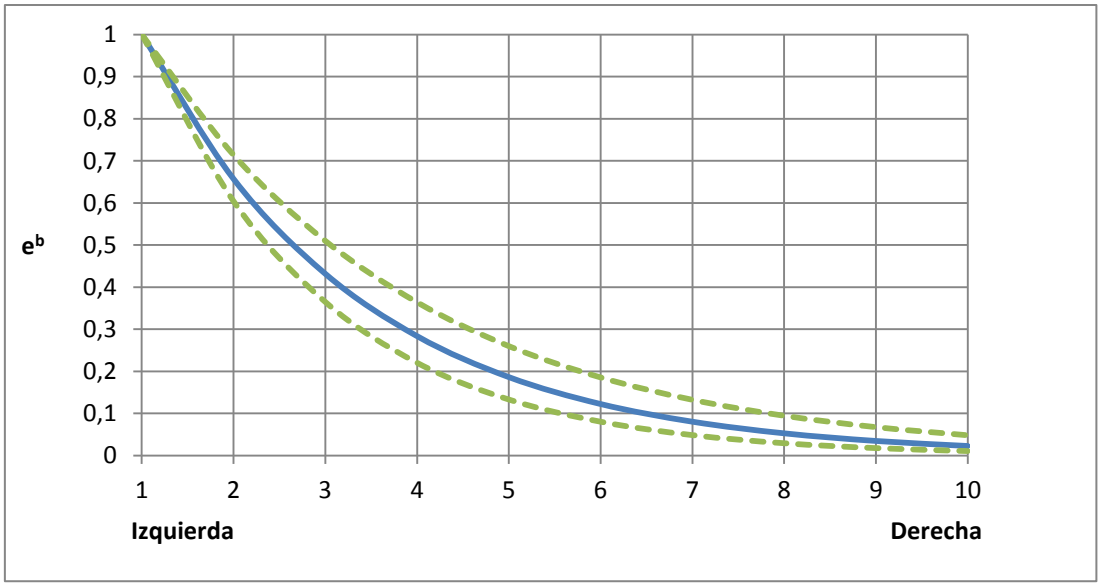
Si bien parecía que las mujeres hacían menos huelga que los hombres o que los jóvenes hacían más huelga que los mayores, era debido a que considerábamos a la población en conjunto. Sin embargo, hay una relación clara entre sexo y ocupación (hay más hombres que trabajan que mujeres) y entre edad y ocupación (hay más jubilados entre los mayores que entre los jóvenes). Una vez que neutralizamos dicho efecto -la relación entre estructura demográfica y actividad- observamos que no hay diferencias en la participación en huelgas. Es decir, la diferencia está en el hecho de que las mujeres

trabajan menos -en figuras estables-, pero, una vez neutralizado dicho efecto, no hay evidencia para afirmar que participan menos en huelgas.

La variable *ideología* sí que tiene relación clara con la participación en huelgas. Como podemos observar, a medida en que la escala se mueve hacia la derecha la participación disminuye en un (1-0,657=34,3%), es decir, por cada "grado" entre izquierda y derecha se pierde la tercera parte de participantes en una huelga. El gráfico siguiente nos ayuda a interpreta el valor de e^b=0,657. Quienes se posicionan con un 2, hacen un 65,67% huelga respecto a quienes se posicionan con un 1, quienes se posicionan con un 3, hacen un 65,7% de huelga respecto a quienes se posicionan con un 2, y un 65,7%x65,7%=43,2% respecto a quienes se posicionan con un 1. Quienes se sitúan en la posición opuesta a la izquierda, los que se han posicionado con un 10 en la escala ideológica, lo hacen $0,657^{10-1}=0,657^9=0,0228=2,3\%$.



El gráfico siguiente permite valorar el intervalo (95%Nc) del coeficiente para la variable Ideología, entre los valores máximo (e^b=0,714) y mínimo (e^b=0,604).



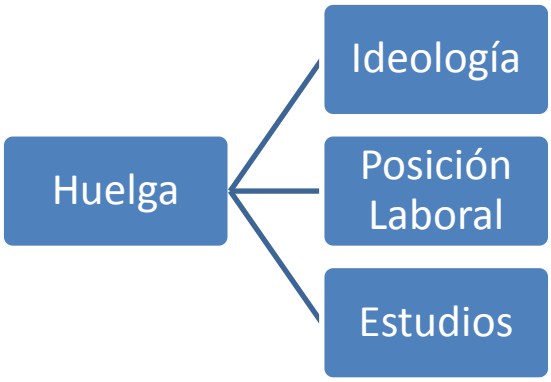
Por su parte, la variable *estatus*, que refleja el componente de clase de los hogares en los que reside el entrevistado, no tiene especial relevancia. Tan sólo podemos apreciar que el grupo de Viejas Clases Medias[15], ofrece diferencias significativas menores del 5%, siendo su logit el menor. Este grupo es el único que se desmarca claramente de la tendencia de huelga, es donde menor proporción de huelguistas encontraremos, una vez neutralizado el efecto del resto de variables.

El *nivel de estudios* tiene un efecto importante y muestra una clara relación positiva entre estudios y huelga, a mayor nivel de estudios crece la participación en huelgas. Por último encontramos que la variable de condición socioeconómica resulta ampliamente significativa. Los obreros cualificados tienen el mayor logit. Esta relación viene a sugerir que la estabilidad permite la realización de huelgas mientras que la precariedad reduce dicha oportunidad.

Una vez realizado un primer análisis, el modelo se ajusta teniendo en cuenta dos criterios. En primer lugar, las variables utilizadas deben ser significativas, y en segundo lugar debe seguirse el *principio de parsimonia*. Este principio señala que ante dos explicaciones posibles hay que optar antes por una explicación sencilla que una compleja.

Atendiendo a ambos principios reduciremos el número de variables. Sexo y edad no tienen capacidad explicativa. Y la variable estatus, además de tener una significación muy parcial, correlaciona con la condición socioeconómica. Por ello, eliminamos estas tres variables. El modelo reducido queda:

[15]El término de "vieja clase media" se refiere a la base patrimonialista –pequeños empresarios- a diferencia de las nuevas clases medias cuya base reside en el salario.



Los coeficientes finales son:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 1[a]	estu			17,456	2	,000			
	estu(1)	,526	,159	10,904	1	,001	1,692	1,238	2,312
	estu(2)	,751	,193	15,095	1	,000	2,120	1,451	3,097
	ideo	-,432	,042	107,694	1	,000	,649	,598	,704
	CONDICION			47,888	9	,000			
	CONDICION(1)	1,399	,331	17,858	1	,000	4,050	2,117	7,747
	CONDICION(2)	1,298	,411	9,987	1	,002	3,664	1,637	8,196
	CONDICION(3)	1,210	,542	4,992	1	,025	3,355	1,160	9,701
	CONDICION(4)	1,285	,346	13,762	1	,000	3,613	1,833	7,122
	CONDICION(5)	1,836	,359	26,192	1	,000	6,274	3,105	12,677
	CONDICION(6)	1,340	,339	15,642	1	,000	3,818	1,966	7,415
	CONDICION(7)	,798	,308	6,725	1	,010	2,220	1,215	4,056
	CONDICION(8)	1,347	,363	13,809	1	,000	3,847	1,890	7,830
	CONDICION(9)	,389	,386	1,015	1	,314	1,476	,692	3,145
	Constante	-,610	,351	3,018	1	,082	,543		

a. Variable(s) introducida(s) en el paso 1: estu, ideo, CONDICION.

Modelo final ajustado.

	b	e^b
ESTUDIOS (Hasta Básicos=0)		
Medios (1)	0,526	1,692**
Universitarios (2)	0,751	2,12**
IDEOLOGÍA	-0,432	0,649**
CONDICIÓN SOCIOECONÓMICA (Directores y Profesionales=0)		
Técnicos y cuadro medios (1)	1,399	4,05**
Pequeños empresarios (2)	1,298	3,664**
Agricultores (3)	1,21	3,355*
Empleados de oficinas y servicios (4)	1,285	3,613**
Obreros cualificados (5)	1,836	6,274**
Obreros no cualificados (6)	1,34	3,818**
Parados (7)	0,798	2,22**
Estudiantes (8)	1,347	3,847**
No clasificables (9)	0,389	1,476
Constante	-0,61	0,543

*Ns>5%
** Ns>1%

El modelo nos indica básicamente que altos estudios y trabajos estables dibujan el perfil de quienes hacen huelga, efecto que crece en la medida en que políticamente se sitúen en el ámbito de la izquierda.

Una vez interpretado el modelo podemos considerar la calidad estadística del mismo:

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1440,246[a]	,139	,201

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Los valores de R^2, después de quitar tres variables, son casi idénticos al modelo con todas las variables. Nuestro modelo es parsimonioso, consigue explicar lo mismo de una forma más sencilla estadísticamente, pero también más clara sociológicamente.

Tabla de clasificación[a]

			Pronosticado		
			huelga		
	Observado		No	Sí	Porcentaje correcto
Paso 1	huelga	No	968	53	94,8
		Sí	279	106	27,5
	Porcentaje global				76,4

a. El valor de corte es ,500

La sensibilidad del modelo es 27,2% y la especificidad 94,8%. Esto quiere decir que nuestro modelo nos orienta para comprender qué variables intervienen en la realización de huelgas, pero, como es casi una norma en investigación social, no tiene capacidad de predicción (sólo conseguiríamos detectar a la cuarta parte de los huelguistas dentro de un grupo de huelguistas). Para avanzar más en la explicación de la participación sería necesario disponer de variables contextuales, variables que generalmente no están incluidas en las encuestas de opinión.

ANEXO

Regresión logística

Notas

Resultados creados		17-MAY-2013 13:59:07
Comentarios		
Entrada	Filtro	dominio = 1 (FILTER)
	Peso	<ninguno>
	Segmentar archivo	<ninguno>
	Núm. de filas del archivo de trabajo	1713
Tratamiento de los datos perdidos	Definición de perdidos	Los valores perdidos definidos por el usuario se consideran como perdidos
Sintaxis		LOGISTIC REGRESSION VARIABLES huelga /METHOD=ENTER estu ideo CONDICION /CONTRAST (CONDICION)=Indicator(1) /CONTRAST (estu)=Indicator(1) /PRINT=CORR SUMMARY CI(95) /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
Recursos	Tiempo de procesador	00:00:00,05
	Tiempo transcurrido	00:00:00,05

Resumen del procesamiento de los casos

Casos no ponderados[a]		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	1406	82,1
	Casos perdidos	307	17,9
	Total	1713	100,0
Casos no seleccionados		0	,0
Total		1713	100,0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

Codificación de la variable dependiente

Valor original	Valor interno
No	0
Sí	1

Codificaciones de variables categóricas

			Codificación de parámetros								
		Frecuencia	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Condición socioeconómica	Directores y profesionales	133	,000	,000	,000	,000	,000	,000	,000	,000	,000
	Técnicos y cuadro medios	151	1,000	,000	,000	,000	,000	,000	,000	,000	,000
	Pequeños empresarios	67	,000	1,000	,000	,000	,000	,000	,000	,000	,000
	Agricultores	25	,000	,000	1,000	,000	,000	,000	,000	,000	,000
	Empleados de oficinas y servicios	118	,000	,000	,000	1,000	,000	,000	,000	,000	,000
	Obreros cualificados	106	,000	,000	,000	,000	1,000	,000	,000	,000	,000
	Obreros no cualificados	146	,000	,000	,000	,000	,000	1,000	,000	,000	,000
	Parados	460	,000	,000	,000	,000	,000	,000	1,000	,000	,000
	Estudiantes	93	,000	,000	,000	,000	,000	,000	,000	1,000	,000
	No clasificables	107	,000	,000	,000	,000	,000	,000	,000	,000	1,000
Nivel de Estudios	Hasta Básicos	515	,000	,000							
	Medios	510	1,000	,000							
	Universitarios	381	,000	1,000							

Bloque 0: Bloque inicial

Tabla de clasificación[a,b]

			Pronosticado		
			huelga		
	Observado		No	Sí	Porcentaje correcto
Paso 0	huelga	No	1021	0	100,0
		Sí	385	0	,0
	Porcentaje global				72,6

a. En el modelo se incluye una constante.

b. El valor de corte es ,500

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-,975	,060	265,933	1	,000	,377

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	estu	23,543	2	,000
		estu(1)	3,186	1	,074
		estu(2)	10,146	1	,001
		ideo	125,962	1	,000
		CONDICION	58,418	9	,000
		CONDICION(1)	17,493	1	,000
		CONDICION(2)	,009	1	,923
		CONDICION(3)	,273	1	,601
		CONDICION(4)	1,505	1	,220
		CONDICION(5)	7,357	1	,007
		CONDICION(6)	3,860	1	,049
		CONDICION(7)	10,123	1	,001
		CONDICION(8)	2,472	1	,116
		CONDICION(9)	6,495	1	,011
	Estadísticos globales		189,906	12	,000

Bloque 1: Método = Introducir

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	210,475	12	,000
	Bloque	210,475	12	,000
	Modelo	210,475	12	,000

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1440,246[a]	,139	,201

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Tabla de clasificación[a]

			Pronosticado		
			huelga		
	Observado		No	Sí	Porcentaje correcto
Paso 1	huelga	No	968	53	94,8
		Sí	279	106	27,5
	Porcentaje global				76,4

a. El valor de corte es ,500

Variables en la ecuación

								I.C. 95% para EXP(B)	
		B	E.T.	Wald	gl	Sig.	Exp(B)	Inferior	Superior
Paso 1[a]	estu			17,456	2	,000			
	estu(1)	,526	,159	10,904	1	,001	1,692	1,238	2,312
	estu(2)	,751	,193	15,095	1	,000	2,120	1,451	3,097
	ideo	-,432	,042	107,694	1	,000	,649	,598	,704
	CONDICION			47,888	9	,000			
	CONDICION(1)	1,399	,331	17,858	1	,000	4,050	2,117	7,747
	CONDICION(2)	1,298	,411	9,987	1	,002	3,664	1,637	8,196
	CONDICION(3)	1,210	,542	4,992	1	,025	3,355	1,160	9,701
	CONDICION(4)	1,285	,346	13,762	1	,000	3,613	1,833	7,122
	CONDICION(5)	1,836	,359	26,192	1	,000	6,274	3,105	12,677
	CONDICION(6)	1,340	,339	15,642	1	,000	3,818	1,966	7,415
	CONDICION(7)	,798	,308	6,725	1	,010	2,220	1,215	4,056
	CONDICION(8)	1,347	,363	13,809	1	,000	3,847	1,890	7,830
	CONDICION(9)	,389	,386	1,015	1	,314	1,476	,692	3,145
	Constante	-,610	,351	3,018	1	,082	,543		

a. Variable(s) introducida(s) en el paso 1: estu, ideo, CONDICION.

Matriz de correlaciones

		Constant	estu (1)	estu (2)	ideo	COND (1)	COND (2)	COND (3)	COND (4)	COND (5)	COND (6)	COND (7)	COND (8)	COND (9)
Paso 1	Constant	1,000	-,299	-,374	-,481	-,614	-,577	-,447	-,643	-,659	-,715	-,800	-,608	-,569
	estu(1)	-,299	1,000	,510	-,001	,004	,062	,009	,005	,061	,054	,101	-,036	-,024
	estu(2)	-,374	,510	1,000	,005	-,125	,177	,121	,095	,237	,192	,214	,089	,041
	ideo	-,481	-,001	,005	1,000	-,004	-,022	,025	-,020	-,044	,028	,042	-,017	,003
	CONDICION(1)	-,614	,004	-,125	-,004	1,000	,533	,403	,646	,604	,648	,717	,613	,584
	CONDICION(2)	-,577	,062	,177	-,022	,533	1,000	,364	,552	,560	,580	,637	,528	,487
	CONDICION(3)	-,447	,009	,121	,025	,403	,364	1,000	,418	,422	,440	,483	,402	,371
	CONDICION(4)	-,643	,005	,095	-,020	,646	,552	,418	1,000	,637	,667	,732	,618	,574
	CONDICION(5)	-,659	,061	,237	-,044	,604	,560	,422	,637	1,000	,672	,737	,611	,561
	CONDICION(6)	-,715	,054	,192	,028	,648	,580	,440	,667	,672	1,000	,771	,639	,590
	CONDICION(7)	-,800	,101	,214	,042	,717	,637	,483	,732	,737	,771	1,000	,699	,648
	CONDICION(8)	-,608	-,036	,089	-,017	,613	,528	,402	,618	,611	,639	,699	1,000	,550
	CONDICION(9)	-,569	-,024	,041	,003	,584	,487	,371	,574	,561	,590	,648	,550	1,000